UiO : Department of Informatics
University of Oslo

Topic Segmentation of Curriculum Vitae

Topic boundary detection

Andreas Bogstad

Master's Thesis Autumn 2015

Topic Segmentation of Curriculum Vitae

Andreas Bogstad

1st December 2015

Abstract

We present an in-depth analysis of Curriculum Vitae documents consisting of unstructured text. We present a collection of Curriculum Vitae Topics with description. We introduce an ontology that gives a formal description of the domain Curriculum Vitae. We presents an analysis that compare the performance of the two PDF extractor algorithms, TIKA and PDFExtract, respectively. We introduce a topic boundary detection algorithm that detects topic boundaries in Curriculum Vitae documents consisting of unstructured text.

To my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 What is a Topic?

According to Reinhart (1981) there is no accepted definition of the notion *topic* in the schools of linguistics. As stated by Gundel and Fretheim (2004) there is a lack of common terminology of the concept *topic* in literature that concerns topic. However, Dijk (1977) puts forward the view that there are two different theoretical notions of *topic*. These notions are *sentence topic* and *discourse topic*, respectively.
The author define the notion sentence topic as what the sentence is about. Furthermore, the author defines discourse topic as what the discourse (text) is about.

Reinhart (1981) understands discourse topics as topics of larger entities that can be more abstract (does not have to) than sentence topics. Moreover, Dijk (1977) states that discourse topics are made of macrostructures. As described in Dijk (1980), the macrostructures (global meanings) are constructed from the sub topics (local meanings) in text by rules that decreases complex information when creating text. Further, the author notes that the discourse topics organises the sub topics in the text. As described in Reynar (1994), the topic (or discourse topic) is usually elaborated by discussion of several sub topics.

1.2 What is Topic Segmentation?

Topic segmentation is the task of dividing a document into meaningful segments. Moreover, (Dias et al., 2007) defines topic segmentation as the task of breaking documents into topically coherent multi-paragraph subparts. Reynar (1998) defines a document as a repository for snippets of written natural language text in any medium which can be accessed, frequently using a computer, after it is created. Karypis and Tagarelli (2008) define a text segment as an indivisible chunk of text, which can in principle be recognized at different levels in the logical structure of the document (e.g. section, paragraph). Moreover, according to Dijk (1980) the text is made of global structures that are called superstructures. Superstructures provides a format (or a schema) that systematise the discourse topics in the text. For

an example, CV documents and scientific papers could be systematised by an established format. Thus, a meaningful segment (topic segment) could be viewed as a segment with a discourse topic that is systematised in a logical structure of the document by a format.
Brown and Yule (1983) suggest that the focus should be on topic shifts rather than attempting to define the notion of topic. Topic shift is a marked point in the text signalising a shift from a topic to another topic. Moreover, the author notes that identifying topic shifts reveals structural information about a text that could be used to divide the text into topic segments. Important, the notion topic shift is used interchangeable with the notion topic boundary. Topic boundary (also called segment boundary) is a boundary that separates two segments.

As described by Riedl and Biemann (2012), there are two forms of topic segmentation, linear topic segmentation and hierarchical topic segmentation, respectively. The authors defines linear topic segmentation as a sequential analysis of topical changes. As described in Choi (2000), the objective in linear topic segmentation is to identify topic boundaries.
Riedl and Biemann (2012)defines hierarchical topic segmentation as finding fine grained subtopic structures in texts.

In this work we will focus on linear topic segmentation.

1.3 Topic Segmentation in Information Retrieval

Identifying document segment is an important problem in Information Retrieval. Finding particular information in a document can be time-consuming and challenging. Searching for information requires processing big documents. Modern information retrieval systems are process-intensive. They use advanced text processing techniques to analyse the document like Part of Speech Tagging, Named Entity Recognition, Wikipedia linking, etc. Often it is required to process the entire document to find the relevant information, usually contained in a small portion of the document. By segmenting the document into segments where each segment is about one topic, these sections can be indexed. Indexed sections make it possible to retrieve information of interest directly after a search engine have found a relevant document. There will be no need to traverse the whole document. Classifying the segments, in a topic hierarchy, will allow search for particular piece of information in the segments with a certain topic and not checking if the whole document in its entirety is about that particular information.

IR systems (that indexing documents consisting of unstructured text) treat a document as being about one topic, but the pure fact is that a document may contain several topics or address subtopics belonging to a discourse topic as described by Reynar (1999). For example in the domain of on-line news, the desired objective is to extract information that describes a news

item from a web page. Such information can be: the author(s) name, the date of the publication, headlines, summary and comments from readers of the news article. This information is not explicitly stated in the web pages. It can appear in the beginning or the end of the web page. Segmenting into sections makes it possible to classify the information that each section contains. This can be done by using a classification system that would for example classify what kind of topic the sections are about. For example in the news domain we can segment a web page that contains an article in to the following topics: headlines, author information, summary, commentary and publication information, and news body. It is also useful to classify the internal structure of the news body. We can separate the background information about an event or entity from the actual news. These issues mentioned above are some of several problems that topic segmentation addresses. The problem of segmenting deals to correctly segmenting a document , that is, that the segmentation creates segments from the document where each of them is just about one topic. Clues from a document can be used to find the most probable topic boundary.

The goal with my proposal is to apply topic segmentation techniques on Curriculum Vitae.

1.4 Topic Segmentation of Curriculum Vitae

To separate CVs into meaningful segments is a challenging task. A CV may be divided into the following segments:

- **Contact information -** segment about the name, address and other contact information of the applicant.
- **Education -** segment about the education level of the applicant. For example, High school, University or College degree courses.
- **Work Experience -** segment about the employment history of the applicant.
- **Skills -** segment about the skills of the applicant. For example, programming skills, languages skills and analytical skills.
- **Publications -** the segment about the published works of the applicant. For example, Article, Dissertation and Book.

Extracting such information from CVs encoded in PDF format (the most common format for CVs) is an extremely difficult task. Different tools exists to convert PDF to text (PDF extractors). For example TIKA[1] uses parser techniques to decode the PDF format. In contrast to this approach, PDFExtract tool (Berg et.al 2012) uses OCR techniques to extract text from the PDF. In both approaches the output of the tools are documents encoded

[1] http://tika.apache.org/

in XML format. The documents consists of a sequence of lines, and they preserves some document structure and layout information from the PDF. However, the *some* document structure and layout information should not be used in a topic segmentation task. The reason for this, is that the structure and layout information in the produced XML documents may variate. Moreover, it is not clear how useful and reliable the structure and layout information could be for segmenting a document. Thus, the XML documents should be treated as unstructured documents. That is, documents which consist of raw text without information about the structure and layout. Unstructured documents creates a serious challenges for topic segmentation. The first challenge is to identify the surface structure of the document. Identifying sections even sentences is a difficult problem. Consider the example in Figure 1.1.

I worked for Apple in the last 10

years.

Education

Finished High School in 2005

Figure 1.1: Example demonstrating the challenge of identifying surface structure in a portion of a unstructured document.

In this example, we have to identify that there are two segments, the first segment is on the first two lines and it speaks about the persons work experience in one sentence.

The topic of the second segment is the education of the applicant which is spread over the last two lines. When reading the original PDF the HR professional can take advantage of page layout, the font shape, text size and blank spaces to determine the segments. The reader can see that *"Education"* has a heading style, for example. In the text extracted from the PDF it can not rely on any other information than the text content. In our example a text segmentation system must conclude that *"Education"* on its own line indicates introduction of a new segment and that the first two lines actually form a sentence talking about *"work experience"* .
Another particularity of the CV domain is that the segments do not follow a specific order. This can be explained by the fact that CVc contains free structure of text. When applying for jobs in the private sector, the *"work experience"* segment can precede the *"education"* in the work oriented CVs. The opposite tends to be true for CVs that are from academic professional, who apply for jobs in the academic sector. These problems make the correct topic segmentation a crucial first step in the analysis of CV information. Having correctly identified segments we can further use information extraction techniques to extract useful data about the applicant and process his/hers application.

In my thesis I will focus on building an approach that address the problem of topic segmentation in unstructured text extracted from CVs encoded in PDF format.

1.5 Overview

Chapter 2 presents relevant works in topic segmentation that describes various clues which are used in identifying topic shifts.

Chapter 3 presents an analysis of the PDF extractor algorithms TIKA and PDFExtract.

Chapter 4 present a collection of Curriculum Vita topics and introduce an ontology that describes the domain Curriculum Vitae.

Chapter 5 describes how the corpus was created, annotated and portioned into data sets. Moreover, an in-depth analysis of CVs is presented and discussed in detail.

Chapter 6 introduce an algorithm that detects boundaries in Curriculum Vitae.

Chapter 7 gives an summary of the conducted work presented in this thesis.

Chapter 8 suggest some future work that can be done.

Chapter 2

Relevant works in Topic Segmentation

Finding potential boundaries is an important step for segmenting a document into meaningful segments. Potential boundary is a term used for indicating that there may be a change of the discourse in a region of text. That is, it signals that at a particular position of the text region, there may be a change of topic.
Locating potential boundaries in a document can be done by finding the placements where regions of text are divided by some sort of markings. As described in Reynar (1999), sentence boundary and paragraph boundary are two examples of markers.
When all markers in a document are identified, potential boundaries can be drawn where these markers occur.

Many approaches in topic segmentation deal with problem of deciding whether a potential boundary is a topic boundary or not. For example, Reynar (1999) views the segmentation of documents as a labelling task. By having a text (from the document that is going to be segmented) and a set of potential boundary placements, the task is either to label these potential boundary placements as *topic boundary* or as *non topic boundary*.

Common strategy of deciding whether a potential boundary is a topic boundary or not is to identify various clues contained in a document. For example, certain words and phrases could signal the end of a topic and the beginning of a new topic.

2.1 Topic Boundary Detection

The literatures about topic segmentation offering various of different ways to attack the topic segmentation problem. The common denominator for these various ways of tackling the topic segmentation problem, is that they all use clues to find topic boundaries. We will give an brief overview of some clues described in the literature about topic segmentation.

2.1.1 Cue words and cue phrases

A stated by Litman (1996) certain words and phrases might be used to unequivocally mark discourse structure in text. That is, a linguistic expression that unequivocally imparts structural information in text.
The literature on phrases and words used as *discourse markers*, are characterized by a lack of uniformity in terminology. Terms that has been coined out are *clue words* (Reichman, 1981), *discourse markers* (Polanyi and Scha, 1984), *cue phrases* (Grosz and Sidner, 1986), *discourse particles* (Schourup, 1985), *rhetorical markers* (Scott and de Souza, 1990), *discourse cues* (Di Eugenio et al., 1997) and *discourse connectives* (Webber et al., 1999). In additional to different terms, the distinction between a word and a phrase are absent. For example, a word and a phrase could both be regarded as a *cue phrase*. Throughout this thesis, the term *cue word* is used and refers to a discourse marker consisting of only one word. Correspondingly, *cue phrase* is used and refers to a discourse marker consisting of at least two words.

As described in Grosz and Sidner (1986), *cue words/phrases* could be used as indicators of topic segment boundaries. That is, a new topic segment may begin since the cue word/phrase indicates a topic shift of the current segment.
According to Reynar (1998), the cue words/phrases are divided into two categories viz. domain-dependentand domain-independent cue words/phrases.

Domain-independent cue words and cue phrases

As noted in Reynar (1998), domain-independent cue words/phrases refers cue words/phrases that could be used to indicate topic shifts in many different domains (or genres). The domain-independent cue words *actually, essentially, otherwise, incidentally, basically, finally and generally* are some examples of domain-independent cue words from (Hirschberg and Litman, 1993) that have been gathered from different sources.
The following shows an example where the domain-independent cue word *actually* signal a topic shift. Consider these two lines: *We went to Paris where we visited the Eiffel tower and the Musée du Louvre. Actually the weather was good that day*. The word *Actually* is a cue word that marks a shift from the topic segment that is about the topic *Sightseeing* to a new topic segment, which is about the topic *Weather*.

Domain-dependent cue words and cue phrases

As described in (Reynar, 1998) domain-dependent cue words/phrase (referred as domain cues) are highly domain-specific. The notion domain-specific is described by the author as the process of manual constructing a list which contains domain-dependent cue words/phrase. Moreover, the list must be constructed before topic segmenting a particular document. The domain-dependent cue words/phrases *hello*, *we'll come back*, *good morn-*

ing, *coming up* and *top stories* are some examples of domain-dependent cue words/phrases from a list (Reynar, 1998) that has been manually constructed, where the domain is broadcasting. The author states that in the domain of broadcasting news, these mentioned cue word/phrase are often used for marking topic shifts from one news segment to another.
The following shows an example where the domain-dependent cue phrase *We'll come back* is used to signal a topic shift. Consider the following lines from a hypothetical news broadcast transcript: *Today it was heavily raining in Oslo. This caused major flooding around our capital. We'll come back on this after our sport news. The World football Championship in Brazil is approaching.* From this example, the cue word/phrase *We'll come back* marks a topic shift from a segment about the topic *weather* to a new segment about the topic *sport*.
As stated by Reynar (1998), the benefit of using domain-dependent cue words/phrases is that the cue words/phrases are credible signals of topic shifts.

Locating cue phrases/words in a text are valuable for topic segmentation of a document.

2.1.2 Word repetition

As described by Hirst and Morris (1991), sentences in a text about a topic have a quality of unity. That is, the sentences in the text about the same topic. This is a property of what the authors refer as cohesion which is described as sentences cohering together to operate as a wholeness. Furthermore, Halliday and Hasan (1976) states that multiple occurrence of phrases and words creates a gives a text a coherence. Moreover, the author pointed out that there is a higher degree of lexical cohesion in a topic segment than an adjacent topic segment. As described in Reynar (1998), the number of occurrence of the same word in an particular topic segment usually are higher than the number of occurrence of the same word occurrence that spans across the topic boundary. Moreover, the author states that few word repetitions that crosses over a potential topic boundary, is a good signal that the potential boundary is a topic boundary.

2.1.3 Character n-gram Repetition

As described by Reynar (1998) using word repetitions to detect a possible topic change, it must be taken into the account that different words may be variants of the same lexeme. A strategy would be to lemmatise the words(to their roots) before looking after repetition. By lemmatising a word to its root form, repetition of same words that are in a singular and plural form can be captured. Consider the following sentence *There are many horses at the farm, the strongest horse is this one*. After lemmatising the words in the sentence, it will be possible to find out that the word *horse* is occurring twice in the sentence.

Obstacles may happen when lemmatising a word that can have two or more different roots. Consider this Norwegian sentence *En fisker fisker fisker* (*A fisher fishes a fish*). The first word is a substantive that refers to a person that is a fisher. The root of this word is *fisker*. The second word is a verb that refers to fishing. The root of this word is *fiske*. The last word refers to the substantive fish, which has the root *fisk*. This example shows that there are three different roots of the word *fisker*. By applying a naive lemmatiser on the verb *fisker*, the lemmatiser may incorrectly lemmatise the word to the root *fisker* or to the root *fisk*. Thus, the relationship between the verb fisker and its root fiske would be lost.
Instead of lemmatising words for then looking after repetition of words, the author suggest that the lemmatising can be dropped, and instead be looking after repetition of character n-grams. For example, consider the sentences *Susanne reports the situation in the Norwegian political landscape. Next week she will be reporting the political landscape of Sweden*. The words *reports* and *reporting* share the common character sequence *report*. Thus, they should be regarded as the same word and the current number of that word is two.

A problem that occurs when identifying repetition of n-grams is that some regular words(e.g., function words) are a substring of longer words. For example, the word *the* is contained in the word *thereafter*. Moreover, if a classification algorithm use character n-gram repetition as a feature, it would identify *the* and *therafter* as repetition of a n-gram *the*. However, the words does not having the same lexeme. Thus, the feature have been assigned an incorrectly value.Therefore, function words (that are short and may be subpart of longer words) should be removed. By doing this, the severity is being decreased but not eliminated. For example, consider the sentence *She shoots a bow and arrow on Friday. On Saturday she is bowling*. The open class word *bow* is in this sentence a subpart of the open class word bowling. However, both of the words have different lexeme. Another example is the sentence *He usually takes the car and visit different lakes in the summer*. The n-gram repetition in this example is the suffix *akes* of the two words that has different lexeme.

2.1.4 Word frequency

As described in Reynar (1998), a better way to detect topic shift than using n-grammar repetition or word repetition is to use word frequency. The reason for this, is that word frequency assume prior knowledge of the frequency of individual words that is in a corpus. Language models are being used to predict the occurrence frequency of words. The advantage of using word frequency above just finding the number of word repetition as an indicator of topic shift ,is that repetitions of words are given different weights based on their likelihood to occur in a document. An example is two neighbouring sentences where each of them contains functions words like 'or', 'and' and 'the'. Despite that there is a repetition of the function words across the two sentence, there is no indication given that the two

sentence belongs to the same topic segment.Another example is when two neighbouring sentences contains the content word anopisthography. The likelihood that these two sentences belongs to the same topic segment is high. Rear words like anopisthography is given more weight than frequent words like 'or', 'and' and 'the'. Thus, repetition of lower frequency words gives a stronger indication of a topic shift than repetition of high frequency words.

2.1.5 Repetition of Named Entities

Name Entities (NEs) are phrases that contain the proper names belonging to name types like *locations, persons and organization*. Consider the following sentences :

```
[ORG NATO] Secretary General [PER Anders Fogh Rasmussen] said
that [ORG NATO] was fully justified in reinforcing the defence
of [LOC Poland] and other [MISC Allies] in wake of the
[LOC Ukraine] crisis.
```

The sentences contains six named entities: NATO is an organization, Anders Fogh Rasmussen is a person, Allies is a miscellaneous name and both Poland and Ukraine are locations.
As described by (Reynar, 1998) the advantage of finding NEs has its origin from the observation that a particular proper name is improbable to arise by chance in topic segments that are neighbours. If a particular proper name is occurring in two or more segments, this will indicate that these segments are about the same topic. Thus, the lines in these segments belong to one topic segment. In contrast, if the particular proper name just occurs in one segment, the proper name cannot be used to indicate that the segment and neighbouring segments are about the same topic.

2.1.6 Introduction of new words

Youmans (1990) states that introduction of new words within a document is often an indication of a topic shift. As noted in Youmans (1995), the frequency of new words have a tendency to increase when a writer introduce a new topic. In contrast, the frequency of new words have tendency to decrease when a writer discusses an old topic.
As described by Reynar (1998), when a new topic segment begins, its discourse normally would contain new places, events and people. That is, new places, events and people that are not contained in the discourse of previous topic segment(s).
The author describes a bias that occur when using introduction of new words as a signal of topic shift. This bias regards the fact that words which are topic independent (e.g., function words) most probable would be used for first time at the beginning of a document. Thus, introduction of new words that are topic independent would incorrectly signal a topic shift. To

address this bias, the topic segmentation of a document should only use new words that are content words as a clue.

Chapter 3

Extracting Raw Text from PDF Documents

In this chapter we provide an analysis of the two PDF extractor algorithms, TIKA and PDFExtract, respectively. The motivation behind this analysis, is that the creation of a corpus requires an algorithm that converts PDF documents into text documents. To establish which of the mentioned algorithms that is best suited for the task, we compared the algorithms performances in extracting text.

3.1 Performance comparison of TIKA and PDFExtract

We had a collection consisting of 1478 CVs encoded in PDF format. The CV documents were written in one of the following languages: Danish(38), Spanish(58), French(697), Italian(5), German(136), Netherlands(4), Portuguese(10), English(495) or Miscellaneous(35).
TIKA and PDFExtract were applied on the collection, and two sets of XML documents were produced (1 from TIKA and 1 from PDFExtract). Moreover, both algorithms failed to extract text from 141 documents in the collection. Consequently, the resulted text documents contained zero characters. These documents were redundant with respect to the comparison of the respective algorithms. Thus, the documents were discarded. Overall, both sets contained each 1337 text documents.
To get an overview of the text quality for text that was extracted by the algorithms, we conducted an analysis of the outputs from the sets.

3.2 Output Analysis

A short inspection of the XML documents (written in English) produced by TIKA and PDFExtract was conducted with purpose of identify possible irregularities.
The most striking irregularity that emerged from the inspection, was that some documents produced by PDFextract contained text lines in incorrect order.

The Figure 3.1 shows a snippets of text from a XML document, where this irregularity is present. In the CV document (PDF format) that the respective XML document's text are extracted from, the words in the first underlined line are all headlines. These headlines are placed in incorrect positions by the algorithm. The second underlined line should have occurred directly after the word *Summary Qualifications*. Moreover, third underlined should have occurred after *Personal Information*. The fourth underlined should have occurred after the *Personal Objective s*. Lastly, the fifth underlined line should have occurred after the word *Education*.

Personal Information: Objective s: Education : Summary Qualifications:
Courses in B. SC (Hons): Father's Name Date of Birth Religion/ Nationality
NIC NO. Domicile To get a challenging job in the field of applied biology
... (text lines)
Genetics/Biochemistry) : (2007-2009) Quaid-i-Azam University Islamabad,

---

Figure 3.1: An example of the incorrect ordering of text produced by PDFExtract.

A possible explanation for this irregularity may be ascribed to what is noted in Berg (2011), that the PDFExtract incorrectly detects columns.
As shown in Table 3.2, the irregularity is absent in the snippets of text from the XML document produced by TIKA.

13503-4815492-4
Objectives:
To get a challenging job in the field of applied biology
(environmental
... (text lines)
M.Phil (Molecular Genetics/Biochemistry) :
2007-2009
() Quaid-i-Azam University Islamabad, Pakistan.
Summary Qualifications:
Courses in B. Sc (Hons):

Figure 3.2: An example correct ordering of text produced by TIKA.

The consequence of improper order of text in documents from a corpus could create complications in an experiment. In detail, incorrect order of text creates noise in a text document. That is, a reduce of the discourse expressed in the PDF document that the text document was converted from. For example, cue words/phrases could occur in text which are in incorrect order. Cue words/phrases are crucial clues that could be used in a topic boundary detection experiment, which examining the context of the document. Thus, misplaced cue words/phrases removes information from text that could be used in detecting topic boundaries. Additionally, misplaced

cue words/phrases could inflict the training of a classification algorithm used in a topic boundary detection experiment, where the focus are on lines individually and not the context (e.g. a misplaced cue word/phrase could incorrectly indicate a false potential topic boundary in a topic segment).
The irregularity does not appear in XML documents produced by TIKA. We wanted to establish whether the performance of TIKA was greater than PDFExtract.

3.3 Statistical Analysis

The performance comparison between TIKA and PDFExtract, was done by comparing the text quality between the text documents from the two sets. We define text quality to be the amount of text extracted (number of characters, tokens and unique tokens) from a PDF document. To establish which of the sets that had highest text quality, a statistical analysis was performed. The statistical analysis consisted of two parts. First, the average amount of tokens, unique tokens and characters in the sets were calculated and compared pairwise(e.g., average number of characters using TIKA compared with PDFExtract). Second, dependent-samples *t* tests were conducted to test whether the highest averages were statistical significant.

To perform the statistical analysis, we created data sets that contained frequency informations (e.g. number of characters) of the text documents from the two sets. In detail, we extracted raw data out of the two sets of XML documents. Furthermore, the raw data were transformed into data sets, which were going to be used in the analyse.
The raw data were created by counting number of character, unique tokens and all tokens that were extracted. As shown in Figure 3.3, the raw data contains frequency informations about two text documents produced by applying TIKA and PDFExtract on one particular PDF document.

/fr/11-0179908_0002.xml
==========
(PDFExtract) (TIKA)
Char 1762 2395
Uniq Tokens 232 291
Total Tokens 349 430
==========

Figure 3.3: An example of raw data that contains frequency informations about two extracted texts produced by applying TIKA and PDFExtract on one particular PDF document.

In this example, the frequency informations provided are number of para-

graphs, characters, unique tokens and tokens extracted only by PDFExtract and not by TIKA (and vice versa).
It is important to note that the extractor algorithms did not succeeded in extracting text from all PDF documents (one algorithm failed and the other succeeded). The failure of extracting text were expressed in the raw data, that all frequency data of the algorithm had the value 0. Furthermore, the failures were not regarded as *missing data*. The failures reflected that the extractor algorithm extracted 0 characters. Thus, producing a XML document with none characters.
Six separated data sets were created by processing the raw data. Moreover, the dataset were divided into two groups, each group linked to the performance of TIKA and PDFExtract, respectively. The data values in the first data set in the two groups, were each the *number of characters* that a particular XML document contained. Correspondingly, the data values in the second and third data sets, were the *number of unique tokens* and *number of all tokens*, respectively.
Important, each row items in the data sets corresponded to frequency information linked to a particular PDF document. Moreover, each row number uniquely identified a PDF document. For example, the row number "5" in the two data sets *number of characters* for TIKA and PDFExtract, are referring to the same PDF document.

From the datasets we calculated (See. Table 3.1) the average number characters, tokens and unique tokens that were extracted by TIKA and PDFExtract.

	Characters	Unique tokens	All tokens
TIKA (mean)	2844	290	524
PDFExtract (mean)	2193	234	394

Table 3.1: The average number of characters, unique tokens and all tokens extracted by TIKA and PDFExtract.

As shown in Table 3.1, all mean scores linked to TIKA are higher than the mean scores of the same variables linked to PDFExtract. Overall, the scores indicates that TIKA has a greater performance than PDFExtract. However, the observed difference between the mean scores could be the results of chance. To establish whether the difference between the mean scores are significant, three dependent matched pair *t* test were conducted.

Dependent matched pair t-test: Number of Characters

A paired-samples t-test was conducted to see whether the difference of the variable *mean characters* from TIKA and PDFExtract was significant. We used the statistical analysis tool SPSS[1] to conduct the dependent matched pair t-test on the two paired datasets *number of characters* linked to TIKA

[1]http://www-01.ibm.com/support/docview.wss?uid=swg27038407

and PDFExtract. The results outputted from the SPSS are shown in Table 3.2 and Table 3.3.

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 TIKA	2844	1337	3817.5	104
PDFExtract	2193	1337	2515	68.78790

Table 3.2: Paired Samples Statistics: Number of characters

	Paired Differences					t	df	sig. (2-tailed)
				99% Confidence Interval of the Difference				
	Mean	Std.Deviation	Std.Error mean	Lower	Upper			
Pair 1 TIKA-PDFExtract	651.7	2475.5	67.7	477.1	826.3	9.6	1336	.000

Table 3.3: Paired Samples Test - Number of characters

There was a significant difference in the scores TIKA (Mean=2844, Std.Deviation=3817.5) and PDFExtract(Mean=2193, Std.Deviation=2515) with the conditions t(1336)=9.6 and p=(0.000). Overall, these results suggests that TIKA is better than PDFExtract in extracting characters from documents encoded in PDF format.

3.3.1 Dependent matched pair t-test: Number of Unique tokens

A paired-samples t-test was conducted to see whether the difference of the variable *mean unique tokens* from TIKA and PDFExtract was significant. From the two paired datasets *number of unique tokens* linked to TIKA and PDFExtract, the SPSS outputted results that are shown in Table 3.4 and Table 3.5.

	Mean	N	Std. Deviation	Std. Error Mean
TIKA	290	1337	259	7
PDFExtract	234	1337	206.5	6

Table 3.4: Paired Samples Statistics - Number of unique tokens

	Paired Differences					t	df	sig. (2-tailed)
				99% Confidence Interval of the Difference				
	Mean	Std.Deviation	Std.Error mean	Lower	Upper			
Pair 1 TIKA-PDFExtract	55.5	159.8	4.4	44.2	66.8	12.7	1336	.000

Table 3.5: Paired Samples Test - Number of unique tokens

There was a significant difference in the scores TIKA (Mean=290, Std.Deviation 259) and PDFExtract(Mean=234, Std.Deviation=206.5) with

conditions t(1336)=12.7 and p=(0.000). Overall, these results suggests that TIKA is better than PDFExtract in extracting unique tokens from documents encoded in PDF format.

3.3.2 Dependent matched pair t-test: Number of all tokens

A paired-samples t-test was conducted to see whether the difference of the variable *mean all tokens* from TIKA and PDFExtract was significant. From the two paired datasets *number of all tokens* linked to TIKA and PDFExtract, the SPSS outputted results that are shown in Table 3.6 and Table 3.7.

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 TIKA	524	1337	754	21
PDFExtract	394.5	1337	456	12.5

Table 3.6: Paired Samples Statistics - Number of all tokens.

	Paired Differences							
				99% Confidence Interval of the Difference				
	Mean	Std.Deviation	Std.Error mean	Lower	Upper	t	df	sig. (2-tailed)
Pair 1 TIKA-PDFExtract	129.8	523	14.3	92.9	166.8	9.1	1336	.000

Table 3.7: Paired Samples Test - Number of all tokens.

There was a significant difference in the scores TIKA (Mean=524, Std.Deviation 754) and PDFExtract (Mean=394.5, Std.Deviation=456) with conditions t(1336)=9.1 and p=(0.000). Overall, these results suggests that TIKA is better than PDFExtract in extracting tokens from documents encoded in PDF format.

3.3.3 Conclusion

The statistical analysis suggests that TIKA extract texts of greater quality than PDF Extract. Moreover, the output analysis discovered that PDFExtract extract text with lines that are in incorrect order in the text document.

Chapter 4

Curriculum Vitae Topics and Ontology

In this chapter we presents a collection of Curriculum Vitae topics with description. Furthermore, we introduce an ontology that describes the domain Curriculum Vitae.

4.1 Collection of Curriculum Vitae Topics

We present a collection of *qualified topics* (systematic entities) that could be found in a CV document. The notion *qualified topic* is defined to be a *discourse topic* that has a tendency to occur in several arbitrary CV documents. The collection is constructed to be generic, that is, the discourse topics that are described in different types of CVs (e.g. Academic CV, Industry CV, Health sector CV and Private sector CV) may be mapped to the topics in this collection. We created the collection by conducting an analysis of 495 English written CV documents encoded in PDF format. Based on the idea that a CV document consists of meaningful segments which are organised in a superstructure (See. Section 1.2), we as human judges, had the task of finding the segments and identify which discourse topic that belonged to the segments. Each identified discourse topic was given a name label. For example, when a text section of a CV was about the topic *skill*, the name label would be skill. From the analysis, we decided which of the identified discourse topics were given the status as a *qualified topic*. The identified *qualified topics* are as follows: Activity, Award, Achievements, Credential, Contact information, Declaration, Education, Experience, Honour, Interest, Miscellaneous, Objective, Professional affiliation, Personal data, Publication, Presentation, Profile, Reference, Summary and Skills. Important, we made one interesting observations during this analysis. The observation was that most of the *discourse topics* that occurred in the CVs were qualified topics. Thus, it may be suggested that CV topics in English written CVs belongs to more or less a static collection of CV topics.
The following gives a detailed description of the CV topics contained in the collection.

Activity
The section in the CV that focus on activities (e.g., hobbies, interests, club participation, and leadership role in an organisation) which highlight the applicants strengths (e.g. leadership and technical skills) belong to the topic *activity*. The following presents an example of a section in the CV which could be mapped to the CV topic activity.

```
Activities       Sigma Nu Fraternity, Delta Alpha Chapter      2003
- 2007
```

- `Vice President, Webmaster`

In this example, the leadership role (Vice President) and a technical responsibility (Webmaster) that the applicant had in an organisation (Sigma Nu) are described.

Award
The section in the CV that list up the awards (e.g. scholarships and employee of the year) that the applicant has received belongs to the topic *award*. The following presents an example of a section in the CV which could be mapped to the CV topic award.

- `Awarded academic scholarship at Amity International School in the year 2005 for meritorious academic performance.`
- `Awarded Rajyapuraskar as a scout for preparing myself for service to God, Country and fellow-men.`

In this example, the scholarship (Amity International School) and the scout emblem (Rajyapuraskar) rewarded to the applicant are described.

Credential
The section(s) in the CV that describes what kind of credentials (e.g. certificate, licensure and patent) that the applicant has received belongs to the topic *credentials*. The following presents an example of a section in the CV which could be mapped to the CV topic credential.

```
OTHER CERTIFICATION

Certification       Institution                        Year
IELTS               Australian Education Organization  2009

Certification       Institution                        Year
Diploma in          United center of languages         2009
language
```

In this example, the certificate (IELTS) with details and the diploma (Diploma in language) with details which are credited to the applicant are described.

Achievement

The section in the CV that list up the applicant's achievements (e.g. developed a data system, discovered a new drug, and previous work promotion) belongs to the topic *achievements*. The following presents an example of a section in the CV which could be mapped to the CV topic achievement.

```
PROJECTS SUCCESSFULLY COMPLETED:

Project on digital notice & announcement board using wireless System,
JNEC, (July 2011-April 2012)
```

In this example, the achievement of a completed project by the applicant is described.

Contact information
The section in the CV that presents contact details (e.g. phone number, email and mail address) that an employer (or other) can use to contact the applicant or a referent belongs to the topic *contact information*. The following presents an example of a section in the CV which could be mapped to the CV topic contact information.

```
   Andreas Bogstad
Lindhaugssvingen 16B
1363 Høvik
+47 45055599
andrebog@student.matnat.uio.no
```

In this example, the person name, mail address (street address, postal code and city), phone number and email address that could be used to contact the applicant are described.

Declaration
The section(s) in the CV that gives a statement declaring truthfulness of the information in the CV, a motto statement, practical information (e.g. when to start in the job, expected salary) by the applicant belong to the CV topic declaration.

```
Declaration

I, Andreas Bogstad, certify that to the best of my knowledge and
belief, these data correctly describes my qualifications, my experience,
and me.
```

In this example, a declaration of the truthfulness of the CV is ibeddescr

Education
The section in the CV that describes the educational background (e.g. University degrees, attended courses and diplomas) of the applicant belongs to the topic *education*. The following presents an example of a section in the

CV which could be mapped to the CV topic education.

```
Studies
20.09.2010 - Start of the studies in the Master of information technologies
Program at the University of Zürich.

09.2002 - 03.2004 Economical Science, University of Neuchatel, 1st
year
```

In this example, details of two University degrees taken (one current taken) by the applicant are described.

Experience

The section(s) in the CV that describes the employment history and other type of experiences, belongs to the topic *experience*. In detail, experience could be professional experience (technical experience in a profession), related work experience (e.g. less technical experience in a profession), volunteer experience (conducted unpaid work), research experience (e.g. participated in research projects) and other work experience.

The following presents and example of a section in the CV which could be mapped to the CV topic experience.

```
PROFESSIONAL EXPERIENCE

Software Engineer IBM, Silicon Valley Lab, CA Feb 2008-Present

RELATED EXPERIENCE Undergraduate Research Assistant 2007
```

- Worked with networking professor and two graduate students for DipZoom project

```
TEACHING EXPERIENCE

Teaching Assistant - Saint Louis University Present
Teaching Assistant - Southern Illinois University 2006-2008
Taught Labs for Calculus 1,2 and 3; Taught College Algebra
```

In this example, details of two University degrees taken (one current taken) by the applicant are described.

In this example, the technical experience in Software Engineering (professional experience), experience of being an Undergraduate Research Assistant (not as technical as experience of being a professor, thus, related experience) and experience in teaching College Algebra (teaching experience), are described.

Honour

The section in the CV that list up the honours that the applicant has received (e.g. honour degree and honour society membership) belongs to the topic *honour*. The following presents an example of a section in the CV which could be mapped to the CV topic honour.

Honours

- *Sloan dissertation fellowship in Mathematics, 1998-99.*
- *Associate of the Academy of Sciences, 2003-2008.*
- *Sloan dissertation fellowship in Mathematics, 1998-99.*

In this example, two honour memberships of the applicant are described.

Interest
The Section(s) in the CV that describes an applicant's interests and activities (does not necessary have to be relevant for the job applied for) that shows some personality characteristics and qualities of the applicant, belongs to the topic *interest*. The following presents an example of a section in the CV which could be mapped to the CV topic interest.

Hobbies

- *Reading new paper*
- *Watching TV*

Field of Interest
Human resource management: To understand how to deal with different kinds of
employees.

In this example, two none job relevant interests (reading newspaper and watching TV) and one job relevant interest (human resource management) of the applicant are described.

Objective
The section in the CV that describes the skills and abilities of the applicant which could benefit the goals of the employer, belongs to the topic *objective*. Moreover, an applicant may write an objective which describes personal and career goals which the applicant expects help from the employer to reach. Thus, description of personal and career goals of the applicant belongs to the topic *objective*.
The following presents an example of a section in the CV which could be mapped to the CV topic objective.

Objective:

To associate with a growth oriented organization where there is opportunity and guidance to develop myself, to face new challenges and to work

```
in an environment where the ideas are encouraged and there is opportunity
for growth and job satisfaction.
```

In this example, personal goal (e.g. develop myself) and career goal (e.g. opportunity for growth) of the applicant are described.

Professional Affiliation
The section in the CV that presents the professional associations that the applicant has a membership in, belongs to the topic *Professional Affiliation*. The motivation to present professional associations in a CV could be to show dedication for a particular profession. Moreover, could show that the applicant is updated about the latest trends in that profession.
The following presents an example of a section in the CV which could be mapped to the CV topic professional affiliation.

```
Professional Memberships

American Economic Association
Association of Christian Economists (founding member)
Institute of Electrical and Electronics Engineers (IEEE)
```

In this example, the two first association are associations for professionals in Economic. Moreover, the last association IEEE is an association for professionals in the area of electrical and electronic engineering.

Personal data
The section(s) in the CV that describes personal data about the applicant (e.g. civil status, age and nationality) belongs to the topic *personal data*.
The following presents an example of a section in the CV which could be mapped to the CV topic objective.

```
Personal Details:

Name:                   Andreas Bogstad
Gender:                 Male
Age:                    28
Nationality:            Norwegian
Citizenships:           Norwegian and Swedish
Civil status:           Married
Children:               Yes
Eye colour:             Blue
```

In this example, personal details about the applicant are described. Note, the line *eye colour: blue* is a biometric data. Biometric data are used when the description of physical appearance is important for the applied job (e.g. Actor).

Publication
The section in the CV that describes the applicant's published work (papers, books and reviews) belongs to the topic *publication*.
The following presents an example of a section in the CV which could be

mapped to the CV topic publication.

```
PUBLICATIONS
Book Chapters

David I. August, Jialu Huangm Thomas B. Jablin, Hanjun Kim, Thomas
R. Mason, Prakash Prabhu, Arun Raman, and Yun Zhang, "Automatic Extraction
of Prallelism from Sequential Code," in Fundamentals of Multicore
Software Development edited by Ali-Reza Adl-Tabatabai, Chapman Hall
/ CRC Press, December 2011.  (ISBN: 978-1439812730)
```

In this example, details of a published book that the applicant have participated in creating are described.

Presentation
The section in the CV that describes the presentations performed by the applicant (e.g. TV interviews and conference presentation) belongs to the the topic *presentation*.
The following presents an example of a section in the CV which could be mapped to the CV topic presentation.

```
INVITED TALKS
```

- Keynote: "Thoughts on Restoring Computing's former Glory,"
  presented at the 2012 Compiler, Architecture and Tools Day,
  Haifa, Israel, November 2012.
- "A Roadmap to Restoring Computing's Former Glory," presented
  at the HiPEAC Industrial Workshop, High-Perfomance and Embedded
  Computing, Charmonix, France, April 2011.

In this example, details of a two presentations conducted by the applicant are described.

Profile
The section in the CV that gives a summary of the applicant's experience, skills, goals and achievements that are relevant for the applied job position, belongs to the topic *profile*.
The following presents an example of a section in the CV which could be mapped to the CV topic profile.

```
PROFILE
A Project Manager / Sr.  Software Engineer.  Experienced in creating
wide range of Web2 ERP based site and portal.  A self trained professional.
Able to work on own initiative as well as a part of a team.  Proved
leadership in managing, developing and motivating teams to achieve
their objectives.  Superior analytical, design and problem solving
skills.  Currently implementing many new technology and applications
like complex architecture of fully dynamic portal with sub site creation
```

```
facility for user from admin.
```
In this example, experience (e.g. creating Web2 ERP based site), skills (e.g. superior analytical) and an upcoming achievement (e.g. implementing fully dynamic portal) are described.

Reference
The section in the CV that gives contact informations of person(s) which can give a comment on the applicant's performance in a past or a current job, belongs to the topic *reference*. Moreover, an applicant may give a statement which says that the contact information is available at request. The statements belongs to the topic *reference*.
The following presents an example of a section in the CV which could be mapped to the CV topic reference.

```
Reference:
Reference will be provided upon request
```

Summary
The section in the CV that gives a summary of the applicant's professional background (e.g. accomplishments, experiences and skills) belongs to the topic *summary*.
The following presents an example of a section in the CV which could be mapped to the CV topic summary.

```
SUMMARY

I have degrees in Electrical Engineering, Computer Science and Mathematics,
and have particular skills in the use of the computer algebra systems
MAGMA and GAP 4, software documentation and the provision of interfaces
to facilitate the accessibility of documentation.  Most recently
I have have developed interactive interfaces to standalone C programs
from within Gab 4.  The results being the two GAP 4 packages ACE
and the soon-to-b-released ANUPQ.
```

In this example, degrees (e.g. software engineering), skills (e.g. manage the computer system MAGMA) and accomplishment (developed interactive interfacer) are described.

Skills
The section(s) in the CV that describes the applicant's skills (e.g. language skills, technical skills, computer skills and personal skills) belongs to the topic *skills*.
The following presents an example of a section in the CV which could be mapped to the CV topic reference.

```
SKILLS

Computer Skills:  MS Office Suite (Word, Excel, Access, PowerPoint,
```

```
FrontPage); Adobe Suite (Photoshop, Illustrator, DreamWeaver, Flash,
Acrobat); Coding:  HTML, CSS, Python, VBA; Share Point, QuarkXpress
Languages:  Fluent in Chinese, proficient in Spanish (reading and
writing), Learning
French
Attributes:  Self confident, Hard work, Ability to work in group
and independently, and Quick Learning.
```

In this example, computer skills (e.g. programming skills in Python), language skills (fluent in Chinese) and personal skills (quick learner) are described.

Miscellaneous

Sections(s) of the CV that cannot be mapped to any of the other topics in this collection of CV topics, belongs to the topic *miscellaneous*

4.2 Curriculum Vitae Topic Ontology

We created an ontology that provides a formal representation of the domain of Curriculum Vitae. In this ontology we defines basic concepts and describe them in detail. Moreover, we define object properties that describes the relationships between the basic concepts.

Basic concepts

:ContactInformation, :PersonalData, :Recognition, :Achievement, :Presentation, :Publication, :Credential,:Declaration :Education, :Objective, :Profile, :Summary, :PersonalDescription, :Reference, :Interest, :Activity, :Affiliation, :Skills, :Location, :Application, :Declaration and :Miscellaneous.

Concept description

:ContactInformation - a concept which relates to the CV topic contact information. This concept is characterised by the following data type properties:

- :full_name - this property relates an instance of this concept to a literal denoting the first name, middle name (if any) and last name of the applicant or referee. The range of the property is either the built-in data type string or Name. An example, :ContactInformation :full_name "Aleksander Jakob Bogstad".

- :first_name - this property relates an instance of this concept to a literal denoting the given name of the applicant or referee. The range

of the property is either the built-in data type string or Name. An example, :ContactInformation :first_name "Aleksander".

- :last_name - this property relates an instance of this concept to a literal denoting the last name of the applicant or referee. The range of the property is either the built-in data type string or Name. An example, :ContactInformation :last_name "Bogstad".

- :middle_name - this property relates an instance of this concept to a literal denoting the middle name of the applicant or referee. The range of the property is either the built-in data type string or Name. An example, :ContactInformation :middle_name "Jakob".

- :email_address - this property relates an instance of this concept to a literal denoting the email address of the applicant or referee. The range of the property is the built-in data type string. An example, :ContactInformation :email_address "jakob-bogstad@astronaut.com".

- :phone_number - this property relates an instance of this concept to a literal denoting a phone number that can be called to communicate with the applicant or referee. The range of the property is the built-in data type string. An example, :ContactInformation :phone_number "004711223344"^^xsd:integer.

- :postal_code - this property relates an instance of this concept to a literal denoting a postal code which is part of a mail address. The range of the property is the built-in data type integer. An example, :ContactInformation :postal_code "1450"^^xsd:integer.

- :street_address - this property relates an instance of this concept to a literal denoting a street address which is part of a mail address. The range of the property is the built-in data type string. An example, :ContactInformation :street_address "8604 Carriage Drive Myrtle Beach, SC 29577".

- :city - this property relates an instance of this concept to a literal denoting a city name which is part of a mail address. The range of the property is the built-in data type string. An example, :ContactInformation :city "Ho Chi Minh City".

- :country - this property relates an instance of this concept to a literal denoting a country which is part of a mail address. The range of the property is the built-in data type string. An example, :ContactInformation :country "Vietnam".

- :state - this property relates an instance of this concept to a literal denoting a state name which is part of a mail address. The range of the property is the built-in data type string. An example, :ContactInformation :state "Texas".

- :region - this property relates an instance of this concept to a literal denoting a region name which is part of a mail address. The range of the property is the built-in data type string. An example, :ContactInformation :region "Burgundy".

:PersonalData - a concept which relates to the CV topic personal data. This concept is characterised with the following data type properties:

- :nationality - this property relates an instance of this concept to a literal denoting the name of the country that the applicant is coming from. The range of the property is either the built-in data type string or Name. An example, :PersonalData :nationality "Sweden".

- :citizenship - this property relates an instance of this concept to a literal denoting the citizenship or citizenships that the applicant have. The range of the property is either the built-in data type string or Name. An example, :PersonalData :citizenship "United States of America & Brazil".

- :country_of_residence - this property relates an instance of this concept to a literal denotion the name of the country that the applicant have been resident for a long period. The range of the property is either the built-in data type string or Name. An example, :PersonalData :country_of_residence "Norway".

- :civil_status - this property relates an instance of this concept to a literal denotion what kind of personal relationship (if any) the applicant has to another person (e.g. married, divorced, civil partnership, widowed and single). The range of the property is the built-in data type string. An example, :PersonalData :civil_status "married".

- :father_name - this property relates an instance of this concept to a literal denotion the name of the applicant's father. The range of the property is either the built-in data type string or Name. An example, :PersonalData :father_name "Andreas Bogstad"

- :mother_name - see father_name . An example, :PersonalData :mother_name "Anna Bogstad".

- :date_of_birth - this property relates an instance of this concept to a literal denoting the birth date of the applicant. The range of the property is the built-in data type date. An example, :PersonalData :date_of_birth "1942-01-08"^^xsd:date.

- :place_of_birth - this property relates an instance of this concept to a literal denoting the name of the place where the applicant was born. The range of the property is either the built-in data type string or Name. An example, :PersonalData :place_of_birth "Ulm, Württemberg, Germany".

- :spouse - this property relates an instance of this concept to a literal denotion the spouse name of the applicant. The range of the property is either the built-in data type string or Name. An example, :PersonalData :spouse "Anna Bella".

- :children - this property relates an instance of this concept to a literal denoting information about the applicant's children(e.g. number of children, name of children, or age of the children). The range of the property is the built-in data type string. An example, :PersonalData :children "Three children". Another example, :PersonalData :children "Name of the children are Huey, Dewey, and Louie"

- :religion - this property relates an instance of this concept to a literal denoting the religious affiliation of the applicant. The range of the property is either the built-in data type string or Name. An example, :PersonalData :religion "Buddhism".

- :appearance - this property relates an instance of this concept to a literal denoting the description of the applicant's physical appearance. The range of the property is the built-in data type string. An example,:PersonalData :appearance "Height 5'7 Hair Brown Eyes Brown".

:Recognition - a concept which relates to the CV topics honour and awards. This concept is characterised with the following data type property:

- :received_recognition - this property relates an instance of this concept to a literal denoting the prices,rewards,honours and awards that the applicant has received. The range of the property is the built-in data type string. An example, :Recognition :received_recognition "Awarded academic scholarship at Amity International School in the year 2005 for meritorious academic performance".

:Achievement - a concept which relates to the CV topic achievement. This concept is characterised with the following data type property:

- :achievement_description - this property relates an instance of this concept to a literal denoting the description of the achievement made by the applicant. The range of the property is the built-in data type string. An example, :Achievement :achievement_description "Winners in intra-college carom competition (2009)".

:Presentation - a concept which relates to the CV topic presentation. This concept is characterised with the following data type property:

- :conducted_presentation - this property relates an instance of this concept to a literal denoting the description of presentation performed by the applicant. The range of the property is the built-in data type string. An example, :Presentation :conducted_presentation "Talks on "Evolutionary relationship of the network by the structural distance computed from the graph Laplacian spoectrum", CMERI-Durgapur, India (8th March,2010)".

:Publication - a concept which relates to the CV topic publication. This concept is characterised with the following data type property:

- :published_work - this property relates an instance of this concept to a literal denoting published work (or soon to be published) by the applicant. The range of the property is the built-in data type string. An example, :Publication :published_work "Economics and the Canadian Economy, with R. Boadway, Canadian Edition. New York: W.W. Norton, 1994.". Another example, :Publication :published_work "2012 "IT IS ALL HERE," Photo Essay, Radical Teacher,#94.University of Illinois Press.".

:Credential - a concept which relates to the CV topic credential. This concept is characterised with the following data type properties:

- :certificate - this property relates an instance of this concept to a literal denoting what kind of certificates, diplomas or licensures that the applicant have been granted. The range of the property is the built-in data type string. An example, :Credential :certificate "Three years National Apprenticeship Certificate in FOOD PRODUCTS from I.T.D.C. from 1989 to 1992".

- :patent - this property relates an instance of this concept to a literal denoting what kind of patents that the applicant have been granted. The range of the property is the built-in data type string. An example, :Credential :patent "Reducing bandwidth requirements for peer-to-peer gaming based on error differences between actual. John R Douceur, Jacob R Lorch, Jeffrey Anson Pang, Frank Christopher Uyeda. U.S. Patent No. 7925601. Apr 12, 2011.".

:Education - a concept which relates to the CV topic education. This concept is characterised with the following data type properties:

- :education_program - this property relates an instance of this concept to a literal denoting the name (or title) of an education program (e.g. course, study program and training) that the applicant has completed. The range of the property is either the built-in data type string or Name. An example, :Education :education_program "MBA in Technology, Market, and Organization from Copenhagen Business School (CBS).". Another example, :Education :education_program "* High school, Electrician".

- :study_sphere - this property relates an instance of this concept to a literal denoting the field, domain or technical area that the applicant's study is addressing towards. The range of the property is either the built-in data type string or Name. An example, :Education :study_sphere "Computer Science".

- :level - this property relates an instance of this concept to a literal denoting the degree level of a study program completed (or going to be completed) by the applicant. The range of the property is either

the built-in data type string or Name. An example, :Education :level "Bachelor of Science". Another example, :Education :level "B.S.".

- :period - this property relates an instance of this concept to a literal denoting the start date and finish date (if any) of a completed study program by the applicant. The range of the property is the built-in data type string. An example, :Education :period "2003-2009".

- :study_arena - this property relates an instance of this concept to a literal denoting the name of the school(e.g. university and company) which provided the education program to the applicant. The range of the property is either the built-in data type string or Name. An example, :Education :study_arena "University of Oslo".

- :description - this property relates an instance of this concept to a literal denoting the full detailed description about a study program completed by the applicant. The range of the property is the built-in data type string. An example, :Education :description "Course for Board members arrange by Vækstfondens, the Danish State investments fund. The series covers the framework for the board of directors, and the expectations it has to work under.". Another example, :Education :description "2003-2009 Ph.D in computer Science, Carnegie Mellon University".

:Objective - a concept which relates to the CV topic objective. This concept is characterised with the following data type property:

- :goal - this property relates an instance of this concept to a literal denoting a relevant goal that the applicant have for the job position. The range of the property is the built-in data type string. An example, :Objective :goal "To work in an environment that would make the best use of my technical knowledge, communication skills and help me groom my technical skills and managerial qualities.".

:Profile - a concept which relates to the CV topic profile. This concept is characterised with the following data type property:

- :goal - See :objective.

:Summary - a concept which relates to the CV topic summary. This concept is characterised with the following data type property:

- :summary_description - this property relates an instance of this concept to a literal denoting the short description of the applicant's professional background. The range of the property is the built-in data type string. An example, :Summary :summary_description "20 years experience, excutive positions in public and private companies – past 14 years in operational P&L management including Sales & Marketing VP, Deputy CEO".

:PersonalDescription - a concept which relates to the CV topics summary, profile or objective. Important, this concept is an alternative to the use of the separated concepts :Summary, :Objective and :Profile. This concept is characterised with the following data type properties:

- :summary_description - see :Summary.
- :goal - see :Objective

:Reference - a concept which relates to the CV topics reference. This concept has the following data type property:

- :reference_description - this property relates an instance of this concept to a literal denoting a comment that states that a reference will be provided upon request. The range of the property is the built-in data type string. An example, :Reference :reference_description "Will be furnished on demand".

:Experience - a concept which relates to the CV topics experience. This concept is characterised by the following data type properties:

- :experience_description - this property relates an instance of this concept to a literal denoting the complete employment information about a certain job that the applicant had(or current have). The range of the property is the built-in data type string. An example, :Experience :experience_description "Professor, July 2012 to Present; Associate Professor, July 2006 to June 2012; Assistant Professor, February 2000 to June 2006; Lecturer, August 1999 to January 2000 Department of Computer Science, Princeton University, Prinston, NJ". Another example, :experience_description :Experience "9.2007-10.2007 Mikron (Boudry): Cabeling of machines' electric boxes during 2 weeks, then job in the field of electronics/industrial machines".
- :employer_name - this property relates an instance of this concept to a literal denoting the name of the entity (e.g. organisation, person and company) that the applicant did (or current) work for. The range of the property is either the built-in data type string or Name. An example, :Experience :employer_name "Princeton University". Another example, :Experience :employer_name "Mikron".
- :period - see :Education. An example, :Experience :period "July 2012 to Present". Another example, :Experience :period "9.2007-10.2007".
- :job_role - this property relates an instance of this concept to a literal denoting the responsibilities or duties of the applicant in a former or current work. The range of the property is the built-in data type string. An example, :Experience :job_role "Cabeling of machines' electric boxes during 2 weeks, then job in the field of electronics/industrial machines"

- :job_title - this property relates an instance of this concept to a literal denoting the name of the job position that the applicant had (or current have). The range of the property is either the built-in data type string or Name. An example, :Experience :job_title "Professor".

:Interest - a concept which relates to the CV topic interest. This interest segment of the CV. This concept is characterised by the following data type property:

- :description_interest - this property relates an instance of this concept to a literal denoting the description of the applicant's interests. The range of the property is the built-in data type string. An example, :Interest :description_interest "The programming, the general culture, the music and the sport".

:Activity - a concept which relates to the CV topic activity. This concept is characterised with the following data type properties:

- :activity_description - this property relates an instance of this concept to a literal denoting the full detailed description about the conducted (or present) activity. The range of the property is the built-in data type string. An example, :Activity :activity_description "Pittsburgh AdFed | October 2012 - Present Volunteer and Participant * Helped facilitate the 2013 Addy's and related events". Another example, :Activity :activity_description "AIGA La Roche Chapter | August 2011 - August 2012 Chapter President * Led La Roche to become an professional student design group".

- :period - see :Education.

- :job_role - see :Experience. An example, :Activity :job_role "Helped facilitate the 2013 Addy's and related events". Another example, :Activity :job_role "Led La Roche to become an professional student design group".

- :activity_arena - this property relates an instance of this concept to a literal denoting the name of the entity (e.g. organisation, club, association) where the applicant had a role, a responsibility or a duty. The range of the property is either the built-in data type string or Name. An example, :Activity :activity_arena "Pittsburgh AdFed".

- :job_title - see :Experience. An example, :Activity :job_title "Present Volunteer and Participant". Another example, :Activity :job_title "Chapter President".

:Affiliation - a concept which relates to the CV topics affiliation. This concept is characterised with the following data type properties:

- :affiliation_description - this property relates an instance of this concept to a literal denoting the complete description of a professional association that the applicant is member of. The range of

the property is the built-in data type string. An example,:Affiliation :affiliation_description "Transatlantic Studies Association (founding member)".

- :affiliation_arena - this property relates an instance of this concept to a literal denoting the name of the association that applicant is member of. The range of the property is either the built-in data type string or Name. An example,:Affiliation :affiliation_arena "Transatlantic Studies Association".

:Skills - a concept which relates to the CV topic skills. This concept is characterised with the following data type properties:

- :computer_skill - this property relates an instance of this concept to a literal denoting the name of the software, hardware, operation systems and programming languages that the applicant have knowledge about. The range of the property is either the built-in data type string or Name. An example, :Skills :computer_skill "Basic experience with Dreamweaver".

- :natural_language - this property relates an instance of this concept to a literal denoting what kind of languages the applicant can speak and write. The range of the property is either the built-in data type string or Name. An example, :Skills :natural_language "French: 10, Czech: 10, German: 4, English: 7 (1 means the worst, 10 means at level "mother tongue", 7: fluently spoken and written at the same level)". Another example, :Skills :natural_language "English language: Good (understanding, writing and speaking)".

- :personal_skill - this property relates an instance of this concept to a literal denoting the personal qualities of the applicant (e.g. communication, team working and organisational skills). The range of the property is the built-in data type string. An example, :Skills personal_skill "* Excellent communication skills, also with peoples other nationalities. * Excellent interpersonal, leadership, and motivational skills.".

- :technical_skill - this property relates an instance of this concept to a literal denoting in-depth knowledge or expertise (e.g. advanced mathematical skills or expert in certain financial instruments) in a particular work area. The range of the property is either the built-in data type string or Name. :Skills :technical_skill "International expertise in online subscription management and Cross Media Communications." Another example, :Skills :technical_skill "Consultations on the media, relaunch, subscriptions, online integration and digitisation."

- :other_skill - this property relates an instance of this concept to a literal denoting all skills that are not technical, personal, natural

language and computer skills. The range of the property is the built-in data type string. An example, :Skills :Other_skill "* Hard worker". Another example, :Skills :Other_skill "Painting (including a number of successful expositions)"

:Declaration - a concept which relates to the CV topic declaration. This concept is characterised with the following data type properties:

- :declaration_description - this property relates an instance of this concept to a literal denoting a statement made by the applicant that the information provided in the CV is correct. The range of the property is the built-in data type string. An example, :Declaration :declaration_description "I, Albert Einstein, certify that to the best of my knowledge and belief, these data correctly describe my qualifications, my experience, and me".

:Miscellaneous - a concept which relates to the CV topics miscellaneous. This concept is characterised with the following data type properties:

- :miscellaneous_description - this property relates an instance of this concept to a literal denoting everything that does not belong to a known topic in the domain of CV. :Miscellaneous :miscellaneous_description " Comany %892 Name Sold".

:Location - a concept which describes a place location. This concept is characterised with the following data type properties:

- :location_name - this property relates an instance of this concept to a literal denoting the name of a geographic location (e.g. region and city). The range of the property is either the built-in data type string or Name. An example, :Location :location_name "San Francisco, CA".

:Applicant - a concept which represents the information about a applicant. This concept is characterised with the following data type property:

- :document_id - this property relates an instance of this concept to literal denoting an identification string that is mapped to a particular CV document. The range of the property is the built-in data type string. An example, :Applicant :document_id "cv_1":

Object property

:hasInterest - the domain and range are: (:Applicant UNION Activity) and (:Interest), respectively.
:hasCredential - the domain and range are: (:Applicant UNION :Education) and (:Credential), respectively.
:hasPublication - the domain and range are: (:Education UNION :Experience UNION :Application) and (:Publication), respectively.
:hasPresented - the domain and range are: (:Applicant) and (:Presentation), respectively.
:hasAchievement - the domain and range are: (:Application UNION :Profile UNION :Experience) and (:Achievement), respectively.

:hasRecognition - the domain and range are: (:Applicant) and (:Recognition), respectively.
:hasAcomplishment - the domain and range are: (:Applicant UNION :Summary) and (:Credential UNION :Publication UNION :Presentation UNION :Achievement UNION :Recognition), respectively.
:hasPersonalEngagement - the domain and range are: (:Applicant) and (:Activity UNION :Affiliation, :Interest), respectively.
:hasContactInformation - the domain and range are: (:Applicant UNION :Reference) and (:ContactInformation), respectively.
:hasPersonalData - the domain and range are: (:Applicant) and (:PersonalData), respectively.
:hasPersonalInformation - the domain and range are: (:Applicant) and (:ContactInformation UNION PersonalData), respectively.
:hasLocation - the domain and range are: (:Education UNION :Experience) and (:Location), respectively.
:hasExperience - the domain and range are: (:Summary UNION :Applicant) and (:Experience), respectively.
:hasTeachingExperience - the domain and range are: (:Applicant) and (:Experience), respectively.
:hasResearchExperience - the domain and range are: (:Applicant) and (:Experience), respectively.
:hasSkills - the domain and range are: (:Experience UNION :Objective UNION :Profile UNION :Summary UNION :Applicant) and (:Skills), respectively.
:hasEducation: - the domain and range are: (:Applicant) and (:Education), respectively.
:hasReference - the domain and range are: (:Applicant) and (:Reference), respectively.
:hasDeclaration - the domain and range are: (:Applicant) and (:Declaration), respectively.
:hasOtherInformation - the domain and range are: (:Applicant) and (:Miscellaneous), respectively.
:

Chapter 5

Corpus Description & Analysis

In this chapter, we provide a description of how the corpus that we used to create our segmentation algorithm was created, annotated and portioned into datasets. Furthermore, we presents an analysis conducted over one of the portioned data sets. The findings from the analysis are discussed in detail.

5.1 Corpus Description

The corpus we used to create our Topic Segmentation algorithm was created in the Sauge Project[1] . The corpus was created by first using TIKA to convert 495 English written CV documents encoded in PDF format, into text documents encoded in XML format. TIKA failed to extract text from 37 CVs. Last step in the corpus creation was to discard 25 documents that contained text of bad quality (e.g. the document contained only one line with few characters). In overall, the created corpus consisted of 433 CV documents encoded in XML format.

The corpus was annotated with the following layers of XML standoff annotation:

- Topic Boundaries and Topic Labels
- Entities and Relations between them found in the CV documents.

We used the topic boundaries in the corpus and created BIO[2] annotation of the documents. In detail, the lines in each of the documents from the corpus were either manual annotated with the tag *B* or *I*. In detail, a line with a topic shift was annotated with the tag *B*. In contrast, a line where the topic shift was absent was annotated with the tag *I*. Moreover, the lines was not annotated with the annotation tag *O*. The mentioned annotation tags corresponds to the **BIO** (**B**egin **I**nside **O**utside) tag format

These tags provides meta information about whether a line is at the beginning (*B*) or inside (*I*) a topic topic segment. Moreover, the lines tagged

[1] http://sauge-project.eu/

[2] See. Ramshaw and Marcus(1995) for more details about **BIO**.

with *B* and *I* belongs to the class *topic boundary* and *non-topic boundary*, respectively.

The Figure 5.1 presents an excerpt from one of the annotated documents in the corpus.

3204 NE 20th Ave B
Portland, Oregon 97212 I
3204 NE 20th Ave I
Portland, 97212 I
Skills/Interests Software Skills B
good with computers I
Languages I
spanish I
Special Skills I
I can sing and I play rugby and soccer. I

Figure 5.1: An example of annotated lines from one of the annotated documents in the corpus.

In this example, the four first lines belongs to a topic segment about the CV topic contact information (See. Section 4.1). The first line in this topic segment ends with the annotation tag *B*. This tag provides the meta information about that the topic segment begins at this line. The rest of the lines in the topic segment ends with the annotation tag *I*. This tag provides the meta information that the lines are inside the topic segment and does not signal topic shifts. Moreover, the six last lines belongs the topic segment about the CV topic skills. The first line in this segment ends with the tag *B* and the rest ends with *I*.

We portioned the annotated corpus into three datasets, training set (259 CVs), developer set (65 CVs) and test set (109 CVs), respectively.

5.2 Analysis

We conducted an analysis of the CVs from the developer set. The motivation of the analyse was to identify patterns and other particularities. This knowledge is essential when developing an algorithm that segments CVs. The following presents and discuss our findings.

5.2.1 Topic shifts

The following presents patterns which could signal topic shifts (See. Section 1.2) in CV documents. Moreover, observations linked to the nature of cue words/phrases are discussed.

Cue words/phrases

As described in the literature (See. Section 2.1.1), cue words/phrases could be used as indicators of topic segment boundaries. Moreover, domain-dependent cue words/phrases are credible signals of topic shifts. This knowledge motivated us to identify cue words/phrases signalising topic shifts. The Table 5.1 shows a list we constructed containing 86 cue words identified in the analysis.

abilities	dear	journal	recognition
academia	declaration	keywords	referees
academics	education	language	reference
academy	employee	languages	references
accomplishments	employment	leadership	research
achievements	exhibitions	license	resume
achievement	experience	matric	rewards
activity	expertise	matriculation	recognition
activities	fellowships	memberships	summary
additional	grants	mission	scholarships
affiliation	highlights	objective	specialities
articles	hobbies	objectives	skill
associations	honor	overview	skills
attributes	honors	passions	skillset
awards	honour	personal	skillsets
biometrics	honours	personals	strength
books	internship	portfolio	strengths
career	internships	professional	talks
capabilities	industry	profile	teaching
certifications	interest	proficiency	technical
competences	interests	presentations	technicality
computer		publication	tutor
contribution		publications	values
curriculum vitae		qualification	websites
cv		qualifications	winnings
			workshop

Table 5.1: List of cue words discovered from the analysis.

The Figure 5.2 show an example of an excerpt of a text document where two cue words from the Table 5.1 signals topic shifts.

have been essential for this element of work. INTERESTS • Keen actor. I participated
in extra-curricular Drama at college. • Enjoy sport. Used to play football for youth team
on Sunday. Publisher • Full, clean driver's licence. REFERENCES Available on request.

Figure 5.2: An example of cue words that signals topic shifts in a snippet of text from a CV document encoded in XML format.

In this example, the cue word *interests* marks a shift from a topic segment about an unknown CV topic (See. Section 4.1) to a new topic segment about the topic *interest*. Furthermore, the cue word *references* marks a shift from the topic segment about *interest* to a new topic segment about *reference*.
Table 5.2 shows some of the identified cue phrases from a list that we constructed, the rest of the identified cue phrases are described in the Appendix (See. Section A.4, A.5 and A.6).

additional information	jobs/employment	programming languages
additional skills	Journal Publications	publications and papers
additional qualifications	key skills	related achievement
activity and honor	language skills	research skills
activities and honors	language and skills	relevant experience
activities/athletics	leaderships experience	relevant skills
awards / achievements	linguistic abilities	seminar attend
awards and honors	management skills	special award/honors/certification
career objectives	mother tongue	special skills
career objective	native tongue	software skills
computer skills	native languag	skills and accomplishment
curriculum vitae	other experience	skills and knowledge
driver's license	other skills	skills & knowledge
employment objective	other employment	technical skill
education and qualification	outside interests	technical skills
experience skills	programming skill	technical knowledge
graduation project done	programming languages	technical qualification
grants and prizes	personal skills	technical proficiencies
grants and awards	personal experience	technical proficien
honors, awards and prizes	personal detail	top skills
honours and awards	personal data	seminar attend
honor and awards	personal information	skills and accomplishment
honor and activity	personal objective	skills and knowledge
highlights of qualification	personale profile	special skills
highlights of profession	professional training	rewards & recognition
it skills	professional accomplishments	software expertise
job experience	professional history	strengths and key skills
job description	programming skill	work experiences

Table 5.2: List of some cue phrases discovered from the analysis.

The Figure 5.3 shows an example of an excerpt of a text document containing three cue phrases from Table 5.2 which signals topic shifts.

PROFESSIONAL SKILLS • Powerful & logical communication ability with planned & standard functional approach • Ability to organize-prioritizes & execute functional responsibilities, a team player. GRADUATION PROJECTS DONE: a.) During my Post Diploma, as a part of course curriculum, have done 2 month project on HR implementation in a Garment unit. b.) And in my Diploma, have done 6 month project on Dyeing of Aloe Vera Fibers and Yarns (Aloe Vera has a functional property of Vitamin E). PROFESSIONAL EXPERIENCE : 1.) Worked as a merchandiser in a Vastra Apparels, Tiruppur from May, 2006-April,2007.

Figure 5.3: An example of cue phrases that signals topic shifts in a snippet of text from a CV document encoded in XML format

In this example, the cue phrase *professional skills* marks a shift from a topic segment to a new topic segment which is about the topic *skill*. Furthermore, the cue phrase *graduation projects done* marks a shift from the topic segment that is about the topic *skill* to a new topic segment which is about the topic *achievement*. Lastly, the cue phrase *professional experience* marks a shift from the topic segment that is about the topic *achievement* to a new topic segment which is about the topic *experience*.

We conducted an analysis of the discovered cue phrases to identify possible patterns. From this analysis we discovered that some cue phrases contained delimiters (',' or '/') and conjunctions ('and' or '&'). Consider the following example of cue phrases in Figure 5.4.

activities and honors
skills and knowledge
activities / athletics
skills & knowledge
activities and honors
honors, awards and prizes
language skills
additional information
special skills

Figure 5.4: Cue phrases with and without delimiter tokens and conjunctions.

As shown in Figure 5.4, the first six cue phrases consists of words which are separated by delimiter tokens or conjunctions. In contrast, the three last cue phrases does not contain delimiter tokens or conjunctions.

This observation suggests that cue phrases could be divided into two

groups: SEPARATOR and NON-SEPARATOR, respectively. SEPARATOR contains cue phrases that have at least one separator. The notion separator is defined to be a token that is a delimiter token or a conjunction. NON-SEPARATOR contains cue phrases without delimiter tokens and conjunctions.
The cue phrases in SEPARATOR had one of the following syntactical forms: *cue word* + separator + *cue word*, *cue word* + separator + noun, noun + separator + *cue word*, noun + separator + noun; and cue phrases in NON-SEPARATOR had one of the syntactical forms: *cue word* + *cue word*, noun + *cue word*, adjective + *cue word*, noun + noun, noun + adjective and noun + prepositional phrase.

Different spellings of cue words/phrases

We observed that certain cue words and words in cue phrases were a variation of a particular lexeme. That is, morphemes with and without inflections. Consider the following cue words:

activity
activities
skill
skills

The cue words *activities* and *activity* are morphemes with without inflection, respectively, both have the same lexeme *activity*. Moreover, skills and skill are with and without inflection and both have the same lexeme *skill*.

In the analysis we discovered that some cue words/phrases were spelled wrong. Consider the following two words:

summary
summari

The word *summary* which could be a cue word, is a correctly spelled word. However, the word *summari* is an incorrect spelling of either the word *summaries or summary*. The origin of the spelling errors could be that the pdf extractor algorithm failed to extract all characters from a particular PDF document. Another explanation may be that the writer of the CV misspelled a word or learned an incorrect word spelling.

The corpus consisted of texts written in one of the following English dialects: British English, American English or World English[1], respectively.
An observation we made was that some American English and British English cue words which had the same meaning, were spelled different. That is, the cue words were different variations of a lexeme. Consider the following cue words: *honor* and *honour*. The cue word *honor* is an American

[1] All the variants of English which are used in countries worldwide.

English word and *honour* is a British English word. They have the same lexeme honour.

Mining cue words/phrases

It is interesting to note that in all cases of this analysis, the cue words/phrases imparted structural information only if the cue words/phrases were a headline in a CV document. This knowledge can be valuable for one important reason. The cue words/phrases can be mined[1] from different sources (e.g. websites) that provides templates or documents that contain headlines. However, mining headlines must be done with some carefulness, since not all headlines marks a topic shift. That is, some headlines may mark the beginning of sub topics. Table A.2 and Table A.3 in the Appendix presents a list of mined cue words and cur phrases, respectively.

Generating new cue words/phrases

We observed that cue words could be part of a cue phrase. For example, the cue phrase *Award and Honors* consist of the two cue words: *Honors* and *Awards*, respectively. Another example is *Other skills*, which consists of one cue word, namely *skills*.
This observation suggest that new cue phrases could be generated by combining two cue words with a separator (e.g. 'and'). Moreover, new cue words could be generated by splitting up cue phrases that belongs to the group SEPARATOR.
For example, the cue words *abilities* and *skills* combined together with the conjunction 'and' generates two new cue phrases *abilities and skills* and *skills and abilities*. The cue words *grants* and *honours* combined together with the conjunction 'and' generates two new cue phrases *grants and honours* and *honours and grants*. Moreover, the cue phrase *reward & recognition* could be split into the two cue words *recognition* and *reward*. The cue phrase *special award/honors/certification* could be split into the cue words *honors*, *certification* and *awards*. Note that the word *special* is an adjective and is not regarded as a cue word. Moreover, the adjective word could be interpreted as either a part of the noun phrase *special award* or part of the noun phrases *special award*, *special honors* and *special certification*. Thus, cue phrases could be split into other cue phrases.

The observation that cue phrases could contain separators enables the possibility to generate new cue phrases. First, the commutativity law of conjunction[2] can be applied on cue phrases with separators. Using this law enables the possibility of change the order of words (or phrases) in a cue phrase with separator. For example, the words in the cue phrase *achievements / awards* could switch position. Thus, generate the new cue phrase *awards / achievements*. Second, the separator (or separators) in a cue phrase could be replaced with an another separator.

[1] Mined is a term used to describe the process of collecting data from difference sources.
[2] See. https://en.wikipedia.org/wiki/Commutativity_of_conjunction

For example, the separator '/' in the cue phrase *activity/athletics* could be replaced by the separator '&'. Thus, generate the new cue phrase *activity & athletics*. Last, cue phrases with more than one separator could be split into smaller cue phrases. For example, the cue phrases *Skills/Interests/Activity* could generate the cue phrases *Skills/Interest*, *Skills/Activity*, *Skills/Interest* and *Interests/Activity*.
A list (not exhaustive) of generated cue phrases can be found in Table A.3 from the Appendix.

5.2.2 Topic continuation

In the analysis we discovered patterns that signalled topic continuations rather than topic shifts. The notion topic continuation is defined to be a point in a text where a topic shift is absence.

Cue words/phrases

We observed that some cue words/phrases only signalised topic continuation in topic segments which were about a CV topic (See. Section 4.1). Moreover, the cue words/phrases signalled subtopic shifts in the segments. Consider the following example shown in Figure 5.5.

---

8 months as operation manager
PERSONAL INFORMATION
Nationality: Pakistan
Resident of: Abbottabad, Pakistan
Birth date: 10 oct 1990
Gender: Male
Marital Status: Unmarried
Number of
Dependants:
Nill
PROFESSIONAL EXPERIENCE

Figure 5.5: Example of cue words/phrases signalising topic continuation and sub topic shifts.

In this example, the cue phrase *PERSONAL INFORMATION* signal the beginning of a new topic segment about the CV topic personal data. The cue words/phrases *Nationality*, *Resident of*, *Birth date*, *Gender*, *Marital Status* and *Dependants* signals topic continuation and sub topic shifts in the topic segment. The cue phrase *PROFESSIONAL EXPERIENCE* signal topic shift to the CV topic experience. Note that all of the cue words/phrases ends with the colon ':'. The mentioned punctuation mark was observed multiple times to occur after a cue word/phrase. The identified cue words/phrases are presented in Table A.7 from the Appendix.

Post-nominal prepositional phrases

In the analysis we discovered lines containing a zero article[1] followed by a prepositional phrase. The phrases are refereed as post-nominal prepositional phrases[2]. Post-nominal prepositional phrases were observed to signal topic shift and topic continuation. The condition for a post-nominal prepositional phrase to signal a topic shift, was when the phrase occurred in the first line of the topic segment. Moreover, the post-nominal prepositional phrase signalled topic continuation when a preceding adjacent line contained a post-nominal prepositional phrase. To establish the first occurrence of post-nominal in a topic segment is a challenging task. In a corpus with blank lines (structural information) a first occurrence could be a post-nominal prepositional phrase occurring after blank lines. This corpus does not contain blank lines.
Figure 5.6 shows an example of post-nominal prepositional phrases which occurs in the same topic segment.

Flat # 116, Westbury Court, Nightingale Lane,
London,
SW4 9AD
048864321792.
abbhfg@hotmail.co.uk
1. **Masters in Business Administration** (International), London-UK
Anglia Ruskin University, Cambridge-UK
2009 to 2010
2. Masters in Computer Sciences, Lahore-Pakistan
Illama Iqbal Open University, Lahore-Pakistan
2000 to 2002
3. Bachelor in Business Administration, London-UK
University of Newcastle, USA
2005 to 2008
4. Bachelor of Arts, Lahore-Pakistan
University of the Punjab, Lahore-Pakistan
1992 to 1994
1. Team Leader, Superdrug Stores Plc.
2004 Till Date
Responsible for store manager regarding day to day retail operations and activities.

Figure 5.6: An example of a postnominal prepositional phrases indicating a new topic and phrases indicating topic continuation.

This example shows a portion of text with three topics: *contact information*,

[1]Zero article is a noun phrase without a determiner
[2]Post-nominal prepositional phrase is a nominal consisting of a *head noun* that is followed by a *prepositional phrase postmodifier*

education and *experience*. All underlined post-nominal phrases are about the same topic *education*. The line containing the bold post-nominal phrase introduces the new topic *education* (topic shift from *contact information* to *education*). The other lines which contains post-nominal phrases indicates a continuation of the topic *education*. Note that the post-nominal phrases are absent from the lines in the topic segment about the topic *experience*(the topic begin in line *1. Team Leader, Superdrug Stores Plc.*).
In the PDF document that this text was extracted from, an empty line was between the topic segments about the topics *contact information* and *education*. The empty lines are often used as structural information to give a signal to the reader of a document about a coming topic shift. This is often the case when the topic shifts from *contact information* to *education* in CV documents.

The Table 5.3 shows a list of post-nominal prepositional phrases that were discovered in the analysis.

areas of association of	assistance in
bachelor of	assisted in
board of	assist in
college of	bachelor degree in
doctorate of	diploma in
design of	degree in
department of	enrolled in
dissertation on	fluent in
highlights of	fluency in
master of	good in
name of	knowledge in
professor of	ms in
place of	mba in
president of	master degree in
resume of	participated in
summary of	proficient in
university of	support in
	science in
	training in
	work in

Table 5.3: List of post-nominal prepositional phrases discovered from the analysis.

Action words

In the analysis we observed that sometimes neighbouring lines in a text document started with an action word[1] (or a period followed by an action word). In addition, the action words had often a suffix *ed* (e.g. Compiled, Designed and Worked). This suggest that the mentioned action words are in past tense[2]. What is interesting to note is that the lines were in the same

[1] Action words are words (usually verbs) that describes an action. For example, the words *write*, *create* and *design* are all action words.

[2] In this section the notion action word refers to an action word in past tense that occur after a period or at the start of a line.

topic segment. Moreover, the action words signalled sub topic shifts in the lines containing them. That is, action words impart structural information in a topic segment. Thus, action words could signal topic continuations.
A possible explanation of the action words imparting structural information, is that the words could be used as bullet points in a CV document. In detail, a text section within a CV which describes actions (e.g. accomplishments) in a structured manner could be viewed as a list. The lines in the section may contain an action word that is used as a bullet point to indicate that a new list item is being described (e.g. describing a new accomplishment). Consider the following example in Figure 5.7.

Team Leader, Superdrug Stores Plc. (2004 Till Date) Key Roles: Responsible for store manager regarding day to day retail operations and activities. Responsible for overseeing the entire store operations. Responsible for personnel management, merchandise selection and presentation and store operations, in the absence of Store Manager. Assisted manager in all areas of resets, remodels, relocation, and new store openings. Assisted manager in interviewing, hiring, training, mentoring, coaching and evaluating performance of hourly associates. Assisted Manager in handling budget planning and tracking. Worked with store managers to ensure merchandise changes aligned with sales patterns. Handled sales tracking, reporting and inventory control.

Figure 5.7: Example of action words that signals topic continuation.

In this example, the action words are underlined. All lines in the text belongs to the topic segment about the CV topic experience. Moreover, the lines where the action words occur describes different accomplishments. Note that the first action word occurs after a cue word (Key Roles:) that signal a topic continuation. This suggest that there exist other conditions that may be included in the understanding of the notion action word.

5.2.3 Repetition

As described in the literature (See Section 2.1.2), that a topic segment may contain repetition of words, motivated us to analyse the corpus to establish possible patterns where repetition of words or phrases are accentuated.

Repetition of Action words

Multiple action words in neighbouring lines in a CV document were sometimes observed. This observation is related to the observation (made on action words in the first analysis) that action words are likely to occur in sections in CV documents that list up points. Since, action words signal a continuation of a topic (as noted in the first analysis about action words), multiple occurrence of action words in neighbouring lines indicates a coherence between the lines. Thus, multiple action words in neighbouring

lines indicates that the respective lines belong to the the same topic segment. Consider the following example in Figure 5.8.

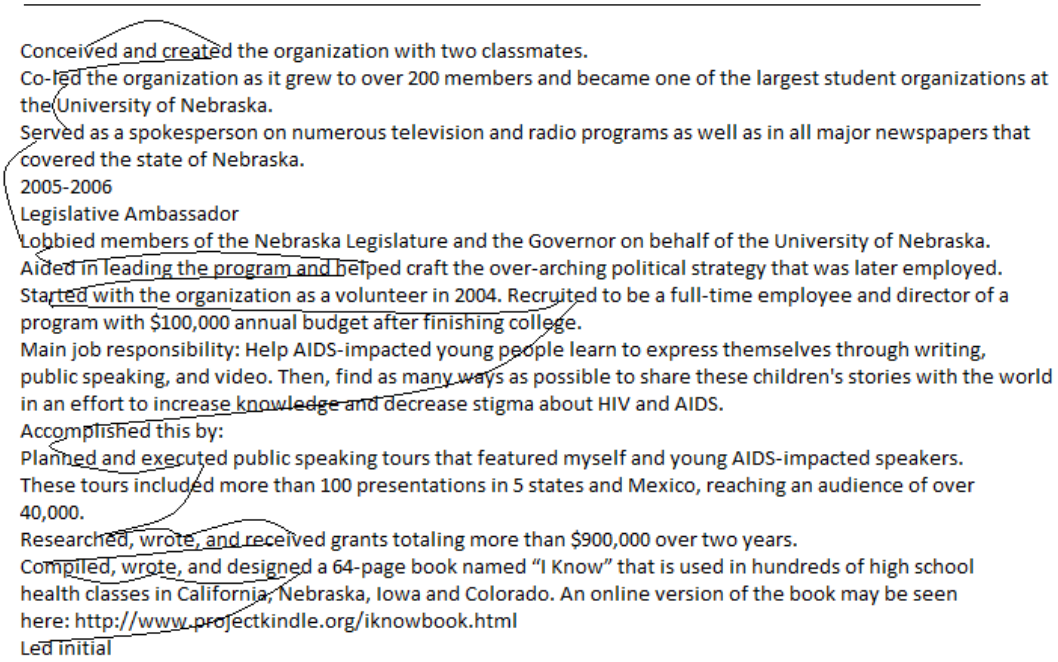

Figure 5.8: Example demonstrating multiple occurrence of action words in a topic segment.

This example shows part of a topic segment that contains the following action words: *Conceived*, *created*, *Co-led*, *Served*, *Lobbied*, *Aided*, *helped*, *Started*, *Recruited*, *Accomplished*, *Planned*, *executed*, *Researched*, *wrote*, *designed*, *Compiled* and *Led*. The drawn lines between the action words visualise the cohesion between the lines.

Repetition of Cue words/phrases

We observed that a topic segments with several lines, could contain multiple cue words/phrases. Moreover, the first cue word/phrase occurrence usually signalled the start of the topic segment (See. Section 5.2.1). The consecutive cue words/phrases usually signalled topic continuation. This observation is related to the observation that cue words/phrases could signal sub topic shifts in a topic segment.
Since, the cue words/phrases imparts structural information about topic continuations and shifts, repetition of cue words/phrases in a topic segment could signal a coherence between lines in a topic segment. Consider the following example in Figure 5.9.

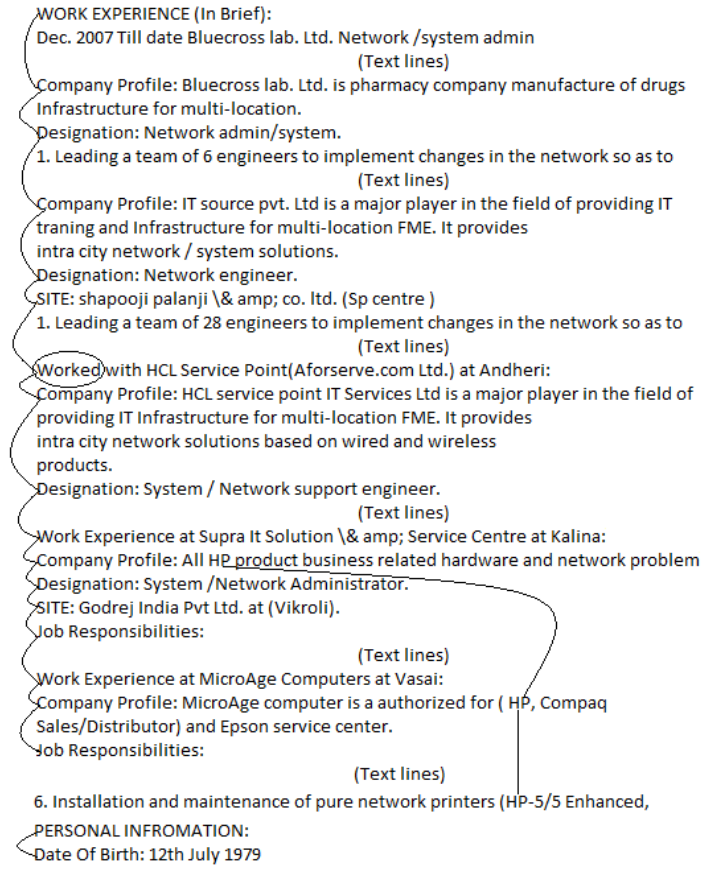

Figure 5.9: Example demonstrating multiple occurrences of cue phrases/words that together creates a cohesion between the lines in a topic segment.

This example shows two topic segments belonging to the CV topics *experience* and *personal data*, respectively. The first topic segment contains the following cue words/phrases: *work experience* (signal a topic shift), *company profile* (signal a topic continuation), *designation* (signal a topic continuation), *site* (signal a topic continuation), *work experience* (signal a topic shift) and *job responsibility* (signal a topic continuation). The second topic segment(the two last lines) contains the two cue phrases *personal information* (signal a topic shift) and *signal topic continuation*. The drawn lines between the cue words/phrases visualises the cohesion between the lines in the respective topic segments.

Note that the action word *worked* is considered as part of the cohesion between the lines. This demonstrates that multiple occurrences of cue words/phrases and action words together creates a cohesion. Furthermore, note the absence of a cue words/phrases in the last line of the first topic segment. The repetition of the proper noun HP could be used to establish that this line is a part of the topic segment. Lastly, note that there are multiple occurrence of the cue phrase *work experience*, where the first occurrence signals topic shift and the other occurrence signal topic continuation. This is

an example of an ambiguity.

Cue word/phrase Ambiguity

An absence of topic shifts in segments containing a cue word were sometimes observed. As noted in Litman and Hirschberg (1987), an ambiguity of cue words appears since cue words can be used in a discourse sense and sentential sense. Similarly, Litman (1996) describes that cue words used in sentential sense imparts semantic information instead of structural information in text. Thus, an absence of the topic shifts could be explained by the ambiguity of cue words. The Figure 5.10 presents an example of cue word ambiguity.

C, Fortran, Perl.
Operating Systems
UNIX System Administrator, DOS, VMS, TSO.
8
Publications
Iversen, Jr, ES, Parmigiani, G, Chen, S (2007). Multiple Model Evaluation Absent the Gold Standard via I-publications
Model Combination. Journal of the American Statistical Association. To Appear.

ancer, European Journal of Human Genetics, IEEE/ACM Transactions on Computational Biology and
Bioinformatics, and Real Estate Economics.
Editorial Board, Medical Decision Making, 12/2003 - 12/2006.
Publications Officer, Risk Section, American Statistical Association, 2006.
Member of NIH Review Panel 'Tumor Microenvironment Network,' September 2006.

Figure 5.10: An example of cue word ambiguity.

In this example, the cue word *Publications* signal the beginning of a new topic segment in the first text excerpt. That is, the cue word imparts structural information. In contrast, the cue word imparts semantic information in the second text excerpt. That is, the cue word *Publication* is a noun which is a part of the noun phrase *Publication Officer* (a job tile), this word does not signal a topic shift.

The analysis showed that cue phrases were less prone to ambiguities than cue words. However, cue phrases which belonged to SEPARATOR (See. Section 5.2.1) had a lower degree of ambiguity in comparison to the cue phrases in NON-SEPARATOR (See. Section 5.2.1). The difference may be

due to the fact that the phrases with separator are less frequently coupled with other terms in a CV, than what is the case for phrases without separators.

We identified another type of ambiguity that was different from the structural and semantic ambiguity. The ambiguity was salient in topic segments containing sub topics with cue words/phrases that usually indicated topic shifts. Consider the following example presented in Figure 5.11.

---

iaison with journalist, authors and festival organizations
SKILLS
COMPUTER SKILLS
MS Office, Adobe Photoshop,Dream Weaver, CMS,HTML, Css Windows, Mac Os,
Mac Software
LANGUAGE SKILLS
Swedish- Native Language. English- Fluent. French- Advanced Conversational

Internship 12 week internship in National Bank of Pakitan city branch
COMPUTER SKILLS
Java, C++, C and Python

Figure 5.11: Example of cue phrase ambiguity when a topic segment contains sub topic segments.

In this example, two excerpts from different CV text documents are presented. The first excerpt shows one cue word and two cue phrases, *skills*, *computer skills* and *language skills*, respectively. The second excerpt show one cue phrase which is *computer skills*.
The ambiguity is that the cue phrase *computer skills* signals topic continuation in the first excerpt and topic shift in the second. In detail, the cue word *skills* in the first excerpts signals a topic shift to the CV topic skills. In contrast, the cue phrases *computer skills* and *language skills* introduces the sub topics, computer skills and language skills, respectively. Moreover, the phrases signals topic continuations rather that topic shifts. However, in the second excerpt the *computer skills* signal a topic shift to the CV topic skills. Thus, the ambiguity is that a cue word/phrase signals the topic shift in one CV document and topic continuation in another.
Note that the cue word *skills* is occurring three times in the first excerpt. The first occurrence is topic shift, the second and third signals topic continuations.

Topic specific words/phrases

As described in the literature (See. Section 2.1.5), a particular proper name is improbable to arise by chance in topic segments that are neighbours. This knowledge motivated us to identify words and phrases that usually only occurs in a topic segment about a particular CV topic (See. Section 4.1).
We observed that certain words may occur together in lines in a topic segment depending on which topic the topic segment was about. These words are categorised as *topic specific words*. The notion *topic specific words* are defined as words that are usually occur together in a topic segment which is affiliated to a specific CV topic. Consider the following example of topic specific words in Table 5.4.

CV topics				
contact information	personal data	honour and award	education	experience
address	name	diploma	bachelor	engineer
telephone	date of birth	honour	master	manager
mobile	nationality	honor	thesis	IEEE
email	place of birth	award	PhD	board
postal code	gender	scholarship	University	organisation
cell phone	citizenship	fellowship	gpa	
e-mail	country of residence	price	college	
residence	religion			
phone	father name			
street	marital status			
tel	place of birth			
email	sex			
	civil status			
	spouse			
	children			

Table 5.4: Example of topic specific words.

In this example, words that usually occur together in one of the following CV topic are presented: *contact information*, *personal data*, *honour*, *award*, *education* or *experience*, respectively. The topic specific words could be used as a signal of cohesion in lines in a topic segment that contains the words. Consider the following example in Figure 5.12.

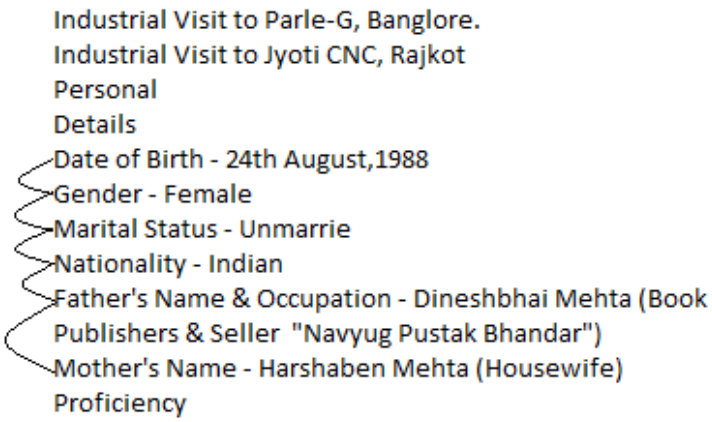

Figure 5.12: Example of group of words that together creates a cohesion between the lines in a topic segment belonging to the CV-topic personal data.

This example shows the following topic specific words/phrases *date of birth*, *gender*, *marital status*, *nationality*, *father name* and *mother name*, respectively. The words creates a cohesion between the lines that contains them (included the line: Publishers & Seller "Navyug Pustak Bhandar")). Moreover, the mentioned lines belongs to the same topic segment about the CV topic personal data. The word proficiency is a cue word which signal a topic shift.

Other patterns

We identified two special characteristics of topic segments about the CV topic contact information.
The topic shift from a topic segment about the topic *contact information* to a new topic segment (e.g. about the topic *education*) is not always marked with a cue word /phrase. This creates a challenging in identifying topic boundaries between the tow topic segments. However, the topic segments about the topic *contact information* possesses two special characteristics. The first characteristic is based on the fact that a CV document usually starts with the contact information on the top of the document. Thus, the first characteristic could be described as follow: the first topic segment to occur in a CV document most likely is a topic segment belonging to the topic *contact information*. The second characteristic is that the topic segment probably contains a line that expresses an email address. Consequently, the line that expresses an email address and all lines above this particular line could be considered to belong to the same topic segment. Moreover, lines below could be considered as part of the same topic segment when the lines expresses phone numbers or when a cue word/phrase occurs in one of the neighbouring lines below.
To identify a line that expresses an email address could be achieved by identify properties of a typical email address. These email address properties are the symbol '@' and the domain part of a email address (e.g. gmail.com, yahoo.com and something.edu). The Figure 5.13, Figure 5.14 and Figure 5.15 present examples of topic segments about the CV topic contact information where the characteristics are prominent.

1361E/11, New Model Town, Ambala road, Kaithal-136027, Haryana, INDIA.
+919467005457
i@tumesh.in
+913509953408
OBJECTIVE

Figure 5.13: Example of topic segment about the topic *contact information*, where a cue word occur below neighbouring lines that expresses an email address and a phone number.

In this example (the personal information has been modified), the lines above i@tumesh.in is a phone number and a postal address. The lines be-

low is a phone number and a cue word. Moreover, the lines that either expresses phone number, postal address or email address all belongs to the same topic segment. The cue word OBJECTIVE indicates the start of a new topic segment about the topic *objective*.

RAHUL SINGH BHAURYAL
RAHUL
SINGH BHAURYAL
+91-9997986858
rsinghbhauryal088@gmail.com
CAREER OBJECTIVE
My hard work , dedication and the ability to acquire new skills will advantage any company

---

Figure 5.14: Example of topic segment about the topic *contact information*, where a cue phrase occur below a line that expresses an email address.

In this example, the line rsinghbhauryal088@gmail.com signal an ending of the topic segment about the topic *contact information* and the cue phrase *CAREER OBJECTIVE* signal the beginning of a new topic segment about the topic *objective*. All lines above the cue phrase *CAREER OBJECTIVE* is an email address, phone number, middle and last name, first name, or full name. Moreover, the lines above belongs to the same topic segment.

Andrea Moreno
Andrea
Moreno
152A Franklin St., Santa Cruz, CA 95060
152A Franklin St., Santa Cruz,
95060
626.340.7862
avalenzu@ucsc.edu
Receptionist, Student Affairs [University of California, Santa Cruz, January2011-Present]

Figure 5.15: Example of topic segment about the topic *contact information*, where a line that belongs to another topic segment occur below a line which expresses an email address.

In this example, the line avalenzu@ucsc.edu signal an ending of the topic about the topic *contact information*. The last line is the first line of a topic segment about the topic *experience*. All lines except the last line belongs to the same topic segment. The lines either expresses an email address, phone number, zip code, postal address, last name, first name, or full name.

We observed that there was minor amount of words in lines with a topic shift than lines with topic continuation. In detail, a CV document usually

use a headlines (cue words/phrases) to introduce a new CV topic in the document. The function of the headlines is to signal topic shifts as clear and descriptive as possible. Consequently, the headlines would not include other words that could reduce the signal of topic shift. Thus, the number of words in lines with a headline which signal a topic shift usually corresponds to the amount of words of the headline. In contrast, the lines with a topic continuation provides more detailed description. Consequently, more words are used then lines with topic shifts. Thus, the number of words in lines with topic continuations are usually greater than with lines with topic shift.
As described in Section 5.2.1, the words in the cue phrases belonged to the following word categories: noun and adjective, respectively. Since, lines usually use cue words/phrases to signal topic shifts, the lines would contain words that are either an adjective or a noun. Furthermore, lines with topic continuation use many words when providing a detailed description. Thus, the words in lines with a topic continuation could belong to other word classes (e.g. determiner) than adjective and noun.

The cue words/phrases and post-nominal prepositional phrases were frequently observed to start at the first position in a lines.

Chapter 6

Boundary Detection in Curriculum Vitae

In this chapter we introduce an algorithm that detects topic boundaries in unstructured text extracted from CVs encoded in PDF format. First we will outline the general architecture. Second, we describe different evaluation methods. Last, we presents, discuss and evaluate experimentations with different configuration of the algorithm.

6.1 General Architecture

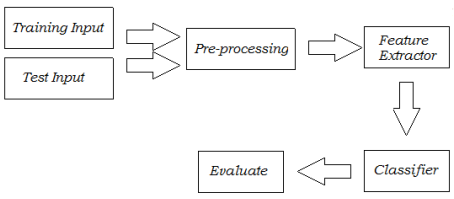

Figure 6.1: The general architecture of the topic boundary detection algorithm

The general architecture of the topic boundary detection algorithm is shown in Figure 6.1. This architecture presents the structure and behaviour of the algorithm. In the following, the first three components of the algorithm and the interaction between them are described.

6.1.1 Input

The input is an BIO tagged line from a document in the training set or test set as described in Chapter 5. Important, the line position and the unique identifier of the document that contains the line are stored in a register. This register provides information about which line positions in particular document has topic boundary.

6.1.2 Pre-processing

The input is pre-processed into a data object that the feature extractor is extracting data from. We use the Java library Apache OpenNLP to conduct some of the pre-processing tasks. As described in OpenNLP (2015), the Apache OpenNLP is a Java library used to process text written in natural language. Moreover, some of the natural languages processing activities that Apache OpenNLP support are tokenization, stemming and Part of Speech (PoS) tagging.
The pre-processing of the input into a data object is described as follow:

1. the last token of the input is an annotation tag. This tag is stored as a data item in the object. The annotation tag is stripped from the input.
2. the stripped line is stored as a data item in the object.
3. the stripped line is tokenized into a list of tokens by a tokenizer provided from the Apache OpenNLP library. Each token (word form) is stored as a data item in the object.
4. the tokens which is an English written word from the list are stemmed by a stemmer from the Apache OpenNLP library. Each stem are stored as data items in the object.
5. the tokens which is an English written word from the list are tagged with a PPoS tag by a Part Of Speech tagger provided from the Apache OpenNLP library.

An example of a data object is shown in Figure 6.2.

String: "Awards & Recognition"
Tag: "B"
Form: ["Awards", "&" ,"Recognitions"]
Stem: ["Award", "&", "Recognition"]
PPoS Tag: ["NNP", "CC", "NNS"]

Figure 6.2: An example of a data object resulted by pre-processing an input.

6.1.3 Feature Extractor

The feature extractor converts the raw data object into a feature vector. This vector contains a set of nominal and quantified features of the line. The following list describes all features in the vector:

Cue word/phrase - Topic shift

- For each token in the line, is the form or the stem of the token a cue word (See. Table 5.1 and A1)?
- For each token in the line, if the form or the stem of the token is a cue word (See. Table 5.1 and A1), what is the position of that token in the line?
- For each cue phrase (See. Table 5.2, A.2, A.3, A.4, A.5 and A.6), is the cue phrase contained in the line?
- For each cue phrase (See. Table 5.2, A.2, A.3, A.4, A.5 and A.6), is the cue phrase contained in the stemmed version of the line?
- For each cue phrase (See. Table 5.2, A.2, A.3, A.4, A.5 and A.6), if the cue phrase is contained in the line, what position in the line does the first word in the cue phrase have?
- For each cue phrase (See. Table 5.2, A.2, A.3, A.4, A.5 and A.6), if the cue phrase is contained in the line, what position in the line does the last word in the cue phrase have?
- For each cue phrase (See. Table 5.2, A.2, A.3, A.4, A.5 and A.6), if the cue phrase is contained in the stemmed version of the line, what position in the line does the first word in the cue phrase have?
- For each cue phrase (See. Table 5.2, A.2, A.3, A.4, A.5 and A.6), if the cue phrase is contained in the stemmed version of the line, what position in the line does the last word in the cue phrase have?

Cue word/phrase - Topic continuation

- For each token in the line, is the form or the stem of the token a cue word (See. A7)?
- For each token in the line, if the form or the stem of the token is a cue word (See. A7), what is the position of that token in the line?
- For each cue phrase (See. A.7), is the cue phrase contained in the line?
- For each cue phrase (See. A.7), is the cue phrase contained in the lemmatized version of the line?
- For each cue phrase (See. A.7), if the cue phrase is contained in the line, what position in the line does the first word in the cue phrase have?

- For each cue phrase (See. A.7), if the cue phrase is contained in the line, what position in the line does the last word in the cue phrase have?
- For each cue phrase (See. A.7), if the cue phrase is contained in the lemmatized version of the line, what position in the line does the first word in the cue phrase have?
- For each cue phrase (See. A.7), if the cue phrase is contained in the lemmatized version of the line, what position in the line does the last word in the cue phrase have?

Action words

- For each token in the line, is the form of the token an action word? (See. A.8)?
- For each token in the line, if the form is an action word (See. A.8), what is the position of that token in the line?

Post-nominal prepositional phrases

- For each post-nominal prepositional phrases (See Table 5.3), is the phrase contained in the line?
- For each post-nominal prepositional phrases (See Table 5.3), if the phrase is contained in the line, what position in the line does the first word in the phrase have?
- For each post-nominal prepositional phrases (See Table 5.3), if the phrase is contained in the line, what position in the line does the last word in the phrase have?

Other features

- What is the total number of tokens in the line?
- For each token in the line, if the token is an English word, what Penn Part of Speech tag (See. A.9) does the token have?
- For each topic specific word/phrase described in Table 5.2.1, does the word/phrase occur in the line?

The features presented are all self explanatory, however some clarifications are needed. The features that has to do with *positions* are numerical features that gives two type of information, binary and positional information, respectively. That is, if a line contains a particular action word, post-nominal prepositional phrase or a cue word/phrase, the feature value would be assigned the position of the word in the line. However, if this is not the case, then the feature value would be assigned the value -1. The idea behind using position as a feature, is explained by the fact that action words, post-nominal prepositional phrase and cue words/phrases may likely to occur in certain positions in the lines than other positions.

Moreover, the positional values provide binary information. That is, if position value is -1, then this states that the line does not contain a action word, post-nominal prepositional phrase, or a cue word/phrase. The opposite is true when the position value is greater that -1.

6.2 Measuring the performance

The following presents evaluation metrics that are widely used to measure the performance of a NLP task.

Recall

As described in Rokach (2010), the recall is a measure that evaluates how good a classification algorithm can detect positive samples. The recall is defined as follows:

$$Recall = \frac{True\ Positve}{True\ Positive + False\ Negative}$$

True Positve is the number of positive samples that were correctly detected. *False Negative* is the number of positive samples that were misclassified as negative samples.

Precision

As described in Rokach (2010), the precision is a measure of the amount of samples classified to the "positive" class that were actual "positive". The precision is defined as follows:

$$Precision = \frac{True\ Positve}{True\ Positive + False\ Positive}$$

False Positive is the number of negative samples that were misclassified as positive samples.

F-score

As described by Resnik and Lin (2010), a great precision could easily be achieved at the expense of recall (and vice versa). This is what Manning and Schütze (1999) notes as a trade of between precision and recall. Further, the authors states that this trade of could make sense in some NLP application (e.g. information retrieval). However, in other NLP applications this trade of does not make sense. In the lather case, the author suggest to use F-score as a measure of the overall performance. As described in Resnik and Lin (2010), this score is a harmonised mean of the two measurements, precision and recall, respectively. The F-score is defined as follows:

$$F_\beta = (\beta^2 + 1) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$$

The parameter β is used to highlight which of the two measurements, that should have a greater influence on the overall measurement. In detail, when the value of the β is lower than 1, then the precision has more influence on the overall measurement. In contrast, when the value is greater than 1, then the recall has more influence on the overall measurement.
The balanced F-score (or F_1-score) is a version of the F-score where both precision and recall has equivalent influence of the overall performance. Moreover, this equivalent influence is expressed by the value of β is set to 1. The formula of the balances F-score is a simplification of the formula for the F_β. The following shows the simplification of the formula F_β into a formula for F_1:

$$F_1 = (1^2 + 1) \times \frac{Precision \times Recall}{1^2 \times Precision + Recall} \Rightarrow 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The mentioned evaluation metrics are used in the experiments presented in the following section.s In the experiments, we are exploring different algorithms abilities to detect topic shifts in lines.

6.3 Experimentation with Classification of Lines

In the following experiments the topic segmentation of a Curriculum Vitae document is viewed as a classification task. That is, by assuming that all lines in a document are potential topic boundaries, the task is to classify each potential topic boundary into one of two classes, *topic boundary* and *non-topic boundary*, respectively.
We wanted to start with simple experimentation with classification of lines by using different established classifiers. Moreover, we wanted to explore whether an ensemble classifier, that combined the different classifiers, would perform better than the best classifier from the same experiment. The classifiers and the ensemble classifier that were used are: Naïve Bayes, C4.5, Random Forest and Stacking, respectively.
We used the java package provided by WEKA to integrate the mentioned classifiers and ensemble classifier into the boundary detection algorithm. As described in (Hall et al., 2009), WEKA is a set of classification algorithms developed for use in data mining puzzles. The algorithms could be integrated inside a Java project.
The following describes the classifiers and the ensemble classifier. Moreover, the philosophy behind them are described.

Naïve Bayes classifier

As described by Rokach (2010), Naïve Bayes is a procedure that utilize a collection of discriminant functions for calculation the likelihood that a particular record pertain to a given class. In detail, given a record, the procedure use Bayes rule to calculate the likelihood of each class where the

features in the feature vector are assumed to be conditional independent of one another. The author explains that since the procedure is founded on the mentioned assumption, the procedure is called Naïve Bayes.

The following presents a definition of Naïve Bayes classifier.

A canonical form of a discriminant function is $g_i(\vec{x}), i = 1, ..., c$. c is number of classes.
A decision rule could be described as chose class ω_i if $g_i(\vec{x}) = max_j g_j(\vec{x})$
Discriminant function is set to be equal the posteriori probability: $g_i(\vec{x})p(\omega_i|\vec{x})$

Bayes rules is as follow

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Naïve Bayes rules can be used to convert posteriori probability into conditional density function connected with a prior probability which is shown as follows.

$$g_i(\vec{x}) = p(\omega_i|\vec{x}) = \frac{p(\vec{x}|\omega_i)p(\omega_i)}{p(\vec{x})}$$

$$p(\vec{x}) = \sum_{j=1}^{c} p(\vec{x}|\omega_j)p(\omega_j)$$

$p(\vec{x})$ is constant for every classes $p(\omega_i)$. Thus, can simplify the discriminant function by removing $p(\vec{x})$. This gives

$$g_i(\vec{x}) = p(\vec{x}|\omega_i)p(\omega_i)$$

Finally, the decision rule of the Naïve Bayes classifiers is as follows
ω_i if $p(\vec{x}|\omega_i)p(\omega_i) = max_j p(\vec{x}|\omega_j)p(\omega_j)$

Decisions Trees

As described by Rokach (2010), a decision tree is a classifier of which the model builds a recursive division of the feature space. Moreover, the model is expressed as a rooted tree. The rooted tree contains three types of nodes, root node, test node, and decision node, respectively. A root node is the first node in the tree, this node has no incoming links. A test node refers to a node with outgoing links and only one incoming link. A decision node refers to a node with no outgoing links and only one incoming link.
The author explains that in a decision tree, every test node use a particular discrete function of feature values in a feature vector to split feature space into at least two sub-spaces. Moreover, every decision nodes are mapped to the most suitable class.

The classification of records is achieved by iterating from the root in the tree to a decision node. Note that the path from the tree node to a decision node is decided by the result of the discrete function of the tree node and each test nodes.
The author notes that building an optimal decision tree from a training set is a challenging process. Moreover, decision tree methods perform only good when the training data set contains few records and that feature vectors contains few features. The tree classifiers C4.5 and Random Forest use heuristic methods to overcome the challenges.

The C4.5 algorithm was developed by Quinlan (1993). As described by (Wu et al., 2007), C4.5 creates decision tree classifiers. The authors describes the algorithm as follow.

Give an collection S containing records, the C4.5 use a divide-and-conquer algorithm to grow a tree. The steps in the algorithm are:

- When all records in S pertain to one particular class or S contains few records, the the tree is decision node which is mapped to the most appropriate class.
- Otherwise, select a test based on feature that has at least 2 outcomes. This test is set to be the root of the tree with one link for every test outcome. Moreover, split S into subsets S_1, S_2... given the outcome of every records. For every subsets, use this method recursively.

To avoid overfitting this tree is pruned by a pruning algorithm. The pruning start from the decision nodes to the root node and is finished when the tree node is reached.

The C4.5 provides a collection of rules where each rule is a conditional statement (e.g. if D and E, then class Y). Moreover, every rules belonging to a class are grouped together. A record is classified using the first rule where the record is satisfying the rule statement. When no rules ha been satisfied, then the record is mapped to a default class.

As defined in Breiman (2001), a random forest is a classifier consisting of a collection of tree structured classifiers $\{h(\mathbf{x}, \Theta_k), k = 1, ...\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$. see Breiman (2001) for more details.

Stacking algorithm

Stacking is an ensemble classifier that were developed by Wolbert (1992). As described in Rokach (2010), the Stacking algorithm use a meta-learning

method to figure out which classifiers that are trustworthy and not trustworthy. Moreover, the author explains that the algorithm is frequently used to combine different classifier.

6.3.1 Experiment with All lines

In this experiment we used a training set and a test set, which were constructed by extracting all lines from the text documents in the training dataset and test data set (See. Chapter 5), respectively. Moreover, the following three classifiers Naive Bayes, C4.5 and Random Forest, and the ensemble classifier Stacking which combined the mentioned classifiers, were trained on training test set with all features described in Section 6.1.3. The trained classifiers and ensemble classifier were applied on the test set and the performance of each was compared to find which one was the best. The Table 6.1 and Table 6.2 shows the results of the classifiers (and the ensemble classifier) in classifying lines into the class *non-topic boundary* and *topic boundary*, respectively.

	Naïve Bayes	Random Forest	C4.5	Stacking
Precision	0.93	0.95	0.95	0.95
Recall	0.97	0.99	0.99	0.99
F_1	0.94	0.97	0.97	0.97

Table 6.1: *Non-topic boundary* results from experiments with all lines. Shows the performance of the classifiers (and ensemble classifier).

As shown in Table 6.1, with F_1-score at 0.94 for Naïve Bayes and 0.97 for the rest of the classifiers (and the ensemble classifier), demonstrates a high performance in classifying lines into the class *non-topic boundary*.
The main goal is to identify topic boundaries. Thus, the rest of the experimentations focus on performance in the classification of lines to the class *non-topic boundary*.

	Naïve Bayes	Random Forest	C4.5	Stacking
Precision	0.46	0.80	0.82	0.87
Recall	0.25	0.46	0.47	0.44
F_1	0.32	0.59	0.60	0.59

Table 6.2: *Topic boundary* results from experiments with all lines. Shows the performance of the classifiers (and ensemble classifier).

As shown in Table 6.2, the results suggest that the Naïve Bayes with a F_1 score at 0.32, performed weaker than the Random Forest, C4.5 and the Stacking with F_1 scores at 0.59, 0.60 and 0.59, respectively. In fact, the precision and recall score of the Random Forest, C4.5 and Stacking are almost double of the scores of the Naïve Bayes.
The classifiers (and ensemble classifier) are trained on an imbalanced training set. This could explain why Naïve Bayes favours towards the majority

class. Thus explaining the under performance of Naïve Bayes.
The C4.5 and Random Forest have high precision but low recall. The algorithms tend to classify line to the class *topic boundary* only if they are quite confident. Furthermore, the Stacking algorithm performs as expected by boosting the precision from the assessments of all three algorithms combined and sacrificing the recall. The stacking algorithm performance is reduced due to the Naïve Bayes under performance. Overall, there is no clear winner between the Random Forest, C4.5 and Stacking, even though the F_1 score of C4.5 is marginal better than that of Random Forest and Stacking. This difference is not significant to conclude that C4.5 is the better than Random Forest.

6.3.2 Experiments with Sampling data

The low recall of the Naïve Bayes, Random Forest, C4.5 and Stacking suggest that the classifiers and ensemble classifier recall are decreased due the fact that the classifiers are trained on an imbalanced data set. In the literature this is known as the *imbalanced data set problem*. The imbalanced data set problem occur when the training set contains samples that belongs to one class which significantly outnumber other samples belonging to an another class. As a consequence, a classifier could be trained to favour the majority class at the expense of the minority class. A training set with all lines from the training dataset described in Chapter) is natural imbalanced. That is, each CV from the training data set contains fewer lines with topic shift than lines with topic continuations.
The observation that the recall of Naïve Bayes, Random Forest, C4.5 and Stacking may be influenced by the imbalanced data set problem, was the motivation of conducting the two following experiments where the classifiers were trained on a balanced training set.

Experiment with Random downsampling

In this experiment we used a balanced training set which was created by using a sample technique known as *Random downsampling* (See. Wang et al., 2011) on all lines from the training data set described in Section 5.1.3. In detail, all lines belonging to the class *topic boundary* and equally number of lines belonging to the class *non-topic boundary* were randomly sampled from the lines in training data set.
All classification algorithms from the previous experiments were trained on the training set and tested on the test data set described in Section 5.1.3. The Table 6.2 shows the results from this experiment.

	Naïve Bayes	Random Forest	C4.5	Stacking
Precision	0.53	0.23	0.34	0.37
Recall	0.52	0.71	0.64	0.64
F_1	0.53	0.41	0.45	0.47

Table 6.3: Result from experiment with Random downsampling

As shown in Table 6.3, the recall of Naïve Bayes, C4.5, Random Forest and Stacking are significant greater than the recall of the same classifiers (and ensemble classifier) from the previous experiment (See. Table 6.1). This suggest that when the mentioned classifiers (and ensemble classifier) are trained on a balanced training set, the classifiers are more inclined at classifying lines with topic shifts than when trained on an imbalanced training set. Furthermore, the F_1 score of the Naïve Bayes is 0.53, which is a greater score than the F_1 score of the same classifier from the previous experiment (See. Table 6.1). This greater performance is a demonstration of our hypothesis, that the classifier perform better on a balanced rather than an imbalanced training set.
Opposite is true for C4.5, Random Forest and Stacking. The F_1 scores at 0.41, 0.45 and 0.47 of the Random Forest, C4.5 and Stacking, respectively, are lower than the F_1 scores of the same classifiers from the previous experiment (See. Table 6.1). This indicates that Random Forest, C4.5 and Stacking, perform weaker when trained on balanced data set than on an imbalanced data set. That is, even though the C4.5 and Random Forest drastically increases the recall, the decision trees are less confident in their decisions which brings the low precision. The algorithms are not capable of learning from the data in the training set. Moreover, the performance of the Stacking is influenced by the low performance of both C4.5 and Random Forest with aspect to precision. The performance of Stacking is greater than the decision trees only because of the Naïve Bayes good performance.

Experiment with Ensemble downsampling

The weak performance of the Random Forest, C4.5 and Stacking from the previous experiment are explained by the low precision of the classifiers and the meta-classifier (See. Table 6.3). A possible explanation of the low precision could be that the randomly selected lines in the training set were not representative of the lines belonging to the class *non-topic boundary*. This motivated us to conduct an experiment without randomly selecting lines that belonged to the class *non-topic boundary*, and consequently discard rest of the lines.

In this experiment we used a sample technique known as *ensemble downsampling* (See. Wang et al., 2011) on all lines in the training data set (See. Chapter 5) to create the training set. In detail, first the training set was divided into two sets, first set (S_1) containing all lines belonging to class *topic*

boundary and second set (S_2) containing all lines belonging to the class *non-topic boundary*. Second, $|S_1|$ of lines were randomly sampled from S_2 into 11 disjoint subsets. Last, all lines from S_1 were added to each subsets. The training set consisted of the 11 subsets, each balanced with equal numbers of positives and negatives.

We created four ensemble classifiers which were as follow: ensemble classifier consisting of 11 Naïve Bayes classifiers, ensemble classifier consisting of 11 Random Forest classifiers, ensemble classifier consisting of 11 C4.5 classifiers, and ensemble classifier consisting of 11 Stacking meta-classifiers where each consisted of Naïve Bayes, Random Forest and C4.5. Furthermore, each classifier in an ensemble classifier were trained on only one of the subsets in the training set. Important, none of the classifiers in an particular ensemble classifier were trained on the same subset. That is, explained with a graph theory analogy, the function between the classifiers in a particular ensemble classifier and the subsets in the training set was bijective.
The decision rule of each ensemble classifiers are described as following:
the probability P of a line belonging to the class *topic boundary* (or *non-topic boundary*), corresponds to the average of the calculated probabilities by classifiers or meta-classifiers for the line belonging to the class. A particular line are classified to the class *topic boundary* when P of line belonging to the class *topic boundary* is greater than the P of belonging to the class *non-topic boundary* (and vice versa).
The ensemble classifiers were applied on the test set (See. Chapter 5) and the results are shown in table 6.4.

	Naïve Bayes	Random Forest	C4.5	Stacking
Precision	0.45	0.33	0.32	0.42
Recall	0.54	0.73	0.69	0.66
F_1	0.49	0.45	0.44	0.52

Table 6.4: Result from experiment with ensemble downsampling.

As shown in Table 6.4, the precision is slightly better from Random Forest and Stacking when compared with the precision scores from the previous experiment (See. Table 6.3). In contrast, the precision precision is slightly lower for Naïve Bayes and C4.5.
Overall, these results suggest that having random extracting of samples belonging to the class *non-topic boundary* was good enough. Utilizing the entire training set did not improve significantly the result when compared to the result in the previous experiment.

6.3.3 Experiment with feature selection

The Naïve Bayes underperformed when trained on an imbalanced training set. A possible explanation could be that the under performance was

a result of the feature combination. This motivated us to conduct an experiment similar to the first experiment, but now using only features that we assessed as strong features. These features were as follows, number of tokens, cue word, cue phrase, action word, and part of speech. The Table 6.5 shows the results.

	Naïve Bayes	Random Forest	C4.5	Stacking
Precision	0.77	0.81	0.83	0.84
Recall	0.37	0.42	0.43	0.47
F_1	0.50	0.55	0.56	0.61

Table 6.5: Result from experiment with feature selection.

As shown in Table 6.5, the F_1 score of Naive is 0.50 and is significant higher than the f-score at 0.32 from the first experiment. This suggest some features caused the Naïve Bayes under performance in the first experiment. As described in Ratanamahatana and Gunopulos (2003), the naïve bayes classier could suffers from being over-sensitive to unimportant. This over-sensitivity could cause a decline of the naïve bayes performance. Moreover, the authors explains that the tree-classifiers are resistant against this over-sensitivity because due to their their split functions. This possible explain that the tree classifiers seems not to be influenced by weak features in the first experiment. The F_1- score of both classifiers are lower than the F_1-scores in the first experiment. A possible explanation for the lower F_1- score could be that some strong features in the first experiment was not included in this experiment.

6.4 Experiment with Conditional Random Fields

The algorithms from the previous experiments did not use information from the context context when classifying lines. This motivated us to explore how the context (all lines in a CV document) could be used to identify topic shifts and topic continuation. In this experiment the topic segmentation of a Curriculum Vitae is viewed as a labelling task. That is, given information from a particular line and from the document that contains the line, the task is to label the line either as *topic boundary* or *non-topic boundary* based on the given informations.
In this experiment we used the sequence labelling algorithm linear-chain Conditional Random Field.

Sutton and McCallum (2010) provides the following definition of linear chain Conditional Random Field:

Let Y, X be random vectors, $\theta = \{\theta_k\} \in \Re^K$ be a parameter vector, and $\{f_k(y_t, y_{t-1}, \mathbf{x}_t)\}_{k=1}^K$ be a set of real-valued feature functions. Then a *linear-*

chain conditional random field is a distribution $p(\mathbf{y}|\mathbf{x})$ that takes the form

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\prod_{t=1}^{T} exp\Big\{\sum_{k=1}^{K}\theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\Big\}$$

Where $Z(\mathbf{x})$ is an instance-spesific normalization function

$$Z(\mathbf{x}) = \sum_{\mathbf{y}}\prod_{t=1}^{T} exp\Big\{\sum_{k=1}^{K}\theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\Big\}$$

We used the Java package provided by Mallet to integrated the CRF into the boundary detection algorithm. As described in McCallum (2002), Mallet is an open source Java based package that provides a collection of sequence tagging algorithms.

The linear-chain Conditional Random Field was trained on the training test set (See. Chapter 5) with all features described in Section 6.1.3. The Table 6.6 shows the results of the CRF labelling lines with the label *topic boundary*, respectively.

	CRF
Precision	0.80
Recall	0.50
F_1	0.62

Table 6.6: Result from experiment with linear-chain Conditional Random Field.

As shown in Table 6.6, the Conditional Random Field performed good. The Conditional Random Field performed good which. The F_1 score was the highest when compared with the result of previous experiment. Overall, this suggest that context contains important information in detecting topic boundaries in a CV document.

6.5 Conclusion

What we learned from the experiment described in Section 6.3.1 are follows:

- C4.5, Random Forest and Stacking had a high performance in classifying lines into the class *non-topic boundary* when they were trained on an imbalanced data set.

- Naïve Bayes performed weaker than C4.5, Random Forest and Stacking.
- Naïve Bayes favours towards majority class when trained on imbalanced training set. This could explain the under performance of the classifier.
- The tree algorithms tends to classify lines to the class *topic boundary* only if they are quite confident.
- The Stacking algorithm algorithm performed as expected by boosting the precision from the assessment of the three algorithms combined and sacrificing the recall. The performance of the Stacking algorithm is reduced due to the Naïve Bayes under performance.
- There was no significant difference in performance between C4.5, Random Forest and stacking.

What we learned from the first experiment described in Section 6.3.2 are as follows:

- The classifiers (and ensemble classifier) are more inclined to classify lines to the class *topic boundary* when they are trained on balanced training set than when trained on an imbalanced training set.
- Our hypothesis that Naïve Bayes perform better on an balanced training set was confirmed.
- The recall of C4.5 and Random Forest is significant higher than recall from the first experiment with imbalanced training set.
- Random Forest, C4.5 and Stacking performed weaker on an balanced training set than on an imbalanced data set. This could be explained with that the C4.5 and Random Forest are less confident in their learning which bring low precision. The algorithms are not capable to learn from the data in the training set.
- The performance of the Stacking algorithm is influenced by the low performance of C.4 and Random Forest with aspect to precision. The algorithm performance is greater than the decision trees only because of the Naïve Bayes good performance.

What we learned from the second experiment described in Section 6.3.2 are as follows:

- That having random extracting of samples belonging to the class *non-topic boundary* was good enough. Utilising the entire training set did not improve significantly the result when compared to the result in the first experiment described in Section 6.3.2.

What we learned from the second experiment described in Section 6.3.3 are as follows:

- Naïve bayes classifier perform better on imbalanced data set with feature selection than without feature selection.

What we learned from the experiment described in Section 6.4 are as follows:

- The Conditional Random Field performed good which. The f_1 score was the highest when compared with the result of previous experiment. This experiment suggests that context contains important information in detecting topic boundaries in a CV document.

Chapter 7

Conclusion

In this work we introduced a topic boundary detection algorithm that detects topic boundaries in unstructured text extracted from CVs encoded in PDF format. The problem of topic segmentation in unstructured text extracted from CVs in PDF format was introduced in Chapter 1. We presented an analysis which compared the performance of the two PDF extractor algorithms TIKA and PDFExtrcat in Chapter 3. Our conclusion based on the results from the analysis was that the TIKA extracts more text than PDFExtract. In chapter 4 we introduced an ontology which gives a formal representation of the domain Curriculum Vita. The concepts described in this ontology were related to the collection of topics that frequently occurred in CVs. This collection was presented in Chapter 4. In chapter 5 we described how the corpus was created, annotated and portioned. In chapter 2 we provided a review of the literature which described patterns that could indicate a topic boundary. We presented an in-depth analysis of a portion of CV documents from the corpus in Chapter 5. In this analysis we presented and discussed patterns that signalled topic shifts and topic continuations. In chapter 6 we introduced the topic boundary detection system. This system was experimented with different configurations and the performances were evaluated with performance measures described in Chapter 6. The result from the experiments were discussed in detail in Chapter 6, and the following describes what we learned from the experimentation:

Experiment with All lines

- C4.5, Random Forest and Stacking had a high performance in classifying lines into the class *non-topic boundary* when they were trained on an imbalanced training set.
- Naïve Bayes performed weaker than C4.5, Random Forest and Stacking. When they all were trained on an imbalanced training set.
- Naïve Bayes favours towards majority class when trained on imbalanced training set. This could explain the under performance of the classifier.

- The tree algorithms tends to classify lines to the class *topic boundary* only if they are quite confident.
- The Stacking algorithm algorithm performed as expected by boosting the precision from the assessment of the three algorithms combined and sacrificing the recall. The performance of the Stacking algorithm is reduced due to the Naïve Bayes under performance.
- There was no significant difference in performance between C4.5, Random Forest and stacking.

Experiment with Random downsampling

- The classifiers (and ensemble classifier) are more inclined to classify lines to the class *topic boundary* when they are trained on balanced training set than when trained on an imbalanced training set.
- Our hypothesis that Naïve Bayes perform better on an balanced training set was confirmed.
- The recall of C4.5 and Random Forest is significant higher than recall from the first experiment with imbalanced training set.
- Random Forest, C4.5 and Stacking performed weaker on an balanced training set than on an imbalanced data set. This could be explained with that the C4.5 and Random Forest are less confident in their learning which bring low precision. The three classifiers are not capable to learn from the data in the training set.
- The performance of the Stacking algorithm is influenced by the low performance of C.4 and Random Forest with aspect to precision. The algorithm performance is greater than the decision trees only because of the Naïve Bayes good performance.

Experiment with Ensemble downsampling

- That having random extracting of samples belonging to the class *non-topic boundary* was good enough. Utilising the entire training set did not improve significantly the result when compared to the result from the experiment *with Random downsampling*.

Experiment with feature selection

- Naïve bayes classifier perform better on imbalanced data set with feature selection than without feature selection.

Experiment with Conditional Random Fields

- The Conditional Random Field performed good which. The F_1 score was the highest when compared with the result of previous experiment. This experiment suggests that context contains important information in detecting topic boundaries in a CV document.

Chapter 8

Future Work

8.1 Structural information

The corpus did not contain blank lines. Blank lines are structural that could be a useful clue to use in a topic boundary detection algorithm. In detail, paragraph tags (or blank lines) could signal a topic shift. Thus, the position in the XML file where a paragraph tag occur could be used as the position of a potential topic boundary.

8.2 Use of the Context

Suggestion to further experimentation is to combine information from the context and lines in an algorithm that detects topic boundaries in Curriculum Vitae.

8.3 System improvement

An improvement of the system would be to use a lemmatizer instead of a stemmer when preprocessing the input of the algorithm. Moreover, programs that could detect spelling error different word spelling could be implemented.

Chapter 9

Bibliography

"Action words". (2015). resumonk. Web. Mars. 2015. <https://www.resumonk.com/resume-builder/resume-keywords-and-action-verbs>.

Brown, G., & Yule, G. (1983). Discourse analysis. Cambridge: Cambridge University Press.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1): 5–32.

Berg, Ø.R., Oepen, S., and Read, J. (2012). Towards High-Quality Text Stream Extraction from PDF. In proceedings of the 50^{nd} Annual Meeting of the Association for Computational Linguistics, pages 98-103 Jeju, The republic of Korea.

Berg, Ø.R. (2011). High precision text extraction from PDF documents.

Choi, F.Y.Y. (2000). Advances in domain independent linear text segmentation. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, pages 26–33, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

Manning, C.D., Schütze. H. (1999). Foundations of statistical natural language processing, MIT Press, Cambridge, MA.

Morris, J. and Hirst, G. (1991).Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17(1), 21–48.

"CV-Headlines". (2015). JobMob. Web. Mars. 2015 .<Web. https://jobmob.co.il/blog/resume-section-headings-titles/>.

Dijk, T.A. (1977). Sentence topic and discourse topic. Papers in Slavic Philology, 1:49-61.

Dijk, T.A. (1980). Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition. Hillsdale, NJ: Erlbaum.

Dias, G., Alves, E., Lopes, J. (2007). Topic segmentation algorithms for text summarization and passages retrieval: an exhaustive evaluation. In: AAAI 2007 Proceedings of the 22nd National Conference on Artificial Intelligence, vol. 2, pp. 1334-1339.

Esser, J. (2006). Presentation in Language: Rethinking Speech and Writing. Tübingen: Gunter Narr Verlag.

Grosz, B. J. and Sidner, C. L. (1986). The structure of discourse. Computational Linguistics, 12(3):175-204.

Gundel, J.K. and Fretheim T. (2004). Topic and Focus. In the Handbook of Pragmatic Theory. Laurence Horn and Gregory Ward (eds.), Blackwell. 174-196.

Halliday, M. A. K., and Hasan, R. (1976).Cohesion in English. London: Longman.

Hirschberg, J., and Litman, D. (1987). Now let's talk about 'now': Identifying cue phrases intonationally. In Proceedings of the Association for Computational Linguistics.

Hirschberg, J., and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. Computational Linguistics, 19(3):501-530.

Hall, M., Frank, E., Holmes, G., Pfahringer B., Reutemann, P., and Witten, I.H. The WEKA data mining software: An update. SIGKDD Explorations, 11(1):10–18, 2009.

Karypis, K. and Tagarelli, A. (2008). S segment-based Approach To Clustering Multi-Topic Documents. Text Mining Workshop, SIAM Datamining Conference, 2008.

Litman, D.J. (1996). Cue phrases classification using machine learning. Journal of Artificial Intelligence Research, 5:53-94. McCallum, A.K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

"OpenNLP". (2015). Apache OpenNLP. Web. <https://opennlp.apache.org/>.

"Penn Treebank Project". (2015). Web. "2015. <http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html>.

Polanyi, L. and Scha, R. (1984). A SYNTACTIC APPROACH TO DISCOURSE SEMANTICS. Proceedings of the 10^{th} International Conference

on Computational Linguistics and 22^{nd} Annual Meeting of the Association for Computational Linguistics, pages 413–419, Stanford, California, USA. Association for Computational Linguistics.

Ratanamahatana, C., Gunopulos, D. (2003). Feature selection for the naive bayesian classifier using decision trees. Applied Artificial Intelligence 17, 475-487.

Reichman, R. (1981). Plain-speaking: A theory and grammar of spontaneous discourse. Ph.D. thesis, Dept. of Computer Science, Harvard University.

Reinhart, T. (1981). PRAGMATICS AND LINGUISTICS: AN ANALYSIS OF SENTENCE TOPICS. Philosophica, 27:53-94.

Reynar, J.C. (1994. 1994. An automatic method of finding topic boundaries. In Proceedings of ACL'94 (Student session).

Reynar, J.C. (1998). Topic segmentation: Algorithms and Applications. Ph.D. thesis, Computer and Information Science, University of Pennsylvania.

Reynar, J.C. (1999). An automatic method of finding topic boundaries. In proceedings of the 37^{nd} Annual Meeting of the Association for Computational Linguistics, pages 357-364, Stroudsburg, PA, USA.

Resnik P, Lin J. Evaluation of NLP systems. In: Clark A, Fox C, Lappin S, editors. The handbook of computational linguistics and natural language processing. Chichester/Malden: Wiley-Blackwell; 2010. p. 271-96.

Riedl, M., Biemann, C. (2012). Text Segmentation with Topic Models. Journal for Language Technology and Computational Linguistics (JLCL), Vol. 27, No. 1, pp. 47–70, August 2012 .

Rokach, L. (2010). Pattern classification using ensemble method. London: World Scientific Pub Co Inc.

Scott, D. and de Souza, C.S. (1990). Getting the Message Across in RST based Text Generation. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation*, Cognitive Science Series. Academic Press.

Eugenio, B.D., Moore, J.D., and Pauolucci M. (1997). Learning features that Predict Cue Usage. In Proceedings of the 35^{th} Annual Meeting of the Association for Computational Linguistics and the 8^{th} Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL97), Madrid.

Scourup, L. (1985). *Common Discourse Particles in English Conversation: Like, Well, Y'know*. New York:Garland.

Sutton, C. and McCallum. (2010). An introduction to conditional random fields. Arxiv preprint arXiv:1011.4088.

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann.

Youmans, G. (1990). Measuring lexical style and competence: the type-token vocabulary curve. Style, 24:584-599.

Youmans, G. (1995). Vocabulary-Management Profiles as Unlabelled Tree Diagrams of Discource. AAAI Technical Report SS-95-06.

Webber, B., Knott, A., Stone, M., and Joshi A. (1999). Discourse Relations: A Structural and Presuppositional Account using Lexicalized TAG. In Meeting of the Association of Computational Linguistics, College Park, MD.

Wang, W., Yaman, S., Precoda, K., Richey, C. and Raymond, G. 2011. Detection of agreement and disagreement in broadcast conversations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT '11, pages 374–378, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., and Motoda, H. (2007). Top 10 algorithms in data mining. Knowledge and information Systems, 14(1), 1-37. doi:10.1007/s10115-007-0114-2.

Appendix A

A.1 Mined cue words

Apprenticeships	Endorsements	Proficiencies
Availability	Exhibits	Recommendations
Accolades	Freelance	Technologies
Courses	Licenses	Testimonials
Credentials	Miscellaneous	Thesis
Conventions	Programs	Theses
Dissertations	Papers	Training
Distinctions		

Table A.1: List of cue phrases mined from the web. Source: "CV-headlines" (2015)

A.2 Mined cue phrases

Academic Background	Military Experience
Academic Experience	Military Background
Academic Honors	Personal Interests
Academic Training	Professional Objective
Additional Experience	Professional Summary
Army Experience	Professional Experience
Areas of Experience	Professional Background
Areas of Expertise	Professional Training
Areas of Knowledge	Professional Affiliations
Athletic Involvement	Professional Associations
Current Research Interests	Professional Memberships
College Activities	Professional Activities
Career Goal	Professional Skills
Career Summary	Professional Publications
Career Related Experience	Related Courses
Career Related Skills	Related Experience
Course Project Experience	Related Course Projects
Community Involvement	Research Grants
Civic Activities	Research Projects
Computer Knowledge	Special Training
Conference Presentations	Specialized Skills
Extra-Curricular Activities	Summary of Qualifications
Educational Background	Social Media Profiles
Educational Qualifications	Technical Experience
Educational Training	Volunteer Work
Education and Training	Volunteer Experience
Employment History	Web Portfolio
Freelance Experience	Writing Samples
Industry Experience	Work History
Internship Experience	Work Experience
Language Competencies and Skills	

Table A.2: List of cue phrases mined from the web. Source: "CV-headlines" (2015)

A.3 Generated cue phrases

activities & athletics	certification and award	honours and prizes	
activities and athletics	certification / award	honors / prizes	
athletics / activities	certification & award	honours / prizes	publications / papers
athletics & activities	certification and honour	honors & prizes	publications & papers
athletics and activities	certification / honour	honours & prizes	papers and publications
awards & achievements	certification & honour	honours, awards and prizes	papers & publications
awards and achievements	certification and honor	honours / awards / prizes	papers / publications
achievements / awards	certification / honor	honours, prizes and awards	qualification and education
achievements and awards	certification & honor	honours / prizes / awards	qualification / education
achievements & awards	certification, award and honor	honors / awards / prizes	qualification & education
awards and honours	certification, award and honour	honors, prizes and awards	recognition and rewards
awards / honours	certification, honor and award	honors / prizes / awards	recognition & rewards
awards & honours	certification, honour and award	honor and certification	recognition / rewards
awards / honors	employment / jobs	honor / certification	rewards and recognition
awards & honors	employment & jobs	honor & certification	rewards / recognition
activity and honour	employment and jobs	honour and certification	rewards & recognition
activity & honour	education / training	honour / certification	skills and languages
activity / honour	education & training	honour & certification	skills & languages
activity & honor	education / qualification	honor, and certification	skills / languages
activity / honor	education / qualification	honour, and certification	skills and languages competencies
activities and honors	grants / awards	honor, certification and award	skills / accomplishment
activities & honors	grants & awards	honour, certification and award	skills & accomplishment
activities / honors	grants and prizes	jobs & employment	skills / knowledge
activities & honours	grants / prizes	jobs / employment	skills and strengths
activities / honours	grants & prizes	knowledge and skills	skills / strengths
accomplishments and skills	honour and activity	knowledge & skills	skills & strengths
accomplishments & skills	honour & activity	knowledge / skills	strengths / key skills
accomplishments / skills	honour / activity	key strengths	strengths & key skills
awards and grants	honor & activity	key strengths and skills	strengths and skills
awards & grants	honor / activity	key strengths / skills	strengths / skills
awards / grants	honors and awards	key strengths & skills	strengths & skills
awards and prizes	honors / awards	key skills and strengths	special award / honor / certification
awards & prizes	honors & awards	key skills / strengths	special award / honour / certification
awards / prizes	honours / awards	key skills & strengths	special award / certification / honor
awards, honors and prizes	honours & awards	languages & skills	special award / certification / honour
awards, prizes and honors	honours and activities	languages / skills	special certification / award / honor
awards, prizes and honours	honours and activities	languages competencies	special certification / award / honour
awards, honours and prizes	honours & activities	prizes and grants	special certification / honor / award
award and certification	honours / activities	prizes & grants	special certification / honour / award
award & certification		prizes / grants	special honor / award / certification
award / certification	honors and activities	prizes, awards and honours	special honour / award / certification
award, certification and honour	honors and activities	prizes, awards and honors	special honor / certification / award
award, honour and certification	honors & activities	prizes, honours and awards	special honour / certification / award
award, certification and honor	honors / activities	prizes, honors and awards	special honor
award, honor and certification	honors and prizes	prizes and honors	special honour
		prizes / honors	special certification
		prizes & honors	special award
		prizes and honours	training and education
		prizes / honours	training / education
		prizes & honours	training & education
		prizes and awards	
		prizes / awards	
		prizes & awards	

Table A.3: List generated of cue phrases.

A.4 Cue phrases I

	bio data	educational qualifications	journal publications
academic / technical profile	career goal	educational record	key achievements
academic achievements	career objective	educational status	key skills
academic achivements	career objectives	educational training	key strengths
academic awards	career related experience	employment experience	language skills
academic background	career related skills	employment history	languages known
academic credential	career summary	employment objective	leaderships experience
academic experience	carrer history	employment record	leisure interests
academic honors	civic activities	executive summary	linguistic abilities
academic prizes	clips/references	experience and highlights	major achievements
academic profile	co-curricular activities	experience gained	major contributions
academic qualification	college activities	experience skills	management skills
academic qualifications	communication skill	extra - curriculum	military background
academic record	community activities	extra circular activity	military experience
academic skills	community based experience	extra curricular activities	mission statement
academic training	community involvement	extra mural activities	mother tongue
academic transcript	computer competences	extra-curricular activities	my motto
activities and awareness	computer knowledge	extracurricular activities	native language
activities and honors	computer literacy	faculty aid award	native tongue
activities and interests	computer proficiency	fellowships and awards	networking activities
activities and organization	computer programs skills	fellowships and grants	other achievements
activities/achievements	computer skill	fellowships and honors	other apprentices
activities/organization	computer skills	film skills	other employment
activity and honor	conference presentations	finance skills	other experience
additional information	conferences attended	foreign languages	other knowledge
additional experience	core skills	foreign languages skills	other languages
additional information	course project experience	formal education	other qualification
additional qualification	covering letter	freelance experience	other pursuits
additional qualifications	current educational status	future achievements	other skills
additional references	current research interests	general qualification	other skils
additional skills	curriculum vitae	graduation project done	outside interests
additional work experience	curriculum vitae/resume	hobbies and activities	past employment
analytical skills	dear madam	honor and activity	patents granted
area of interest	dear sir / madam	honor and awards	personal abilities
areas of experience	dears sir	honors	personal attributes
areas of expertise	driver license	honors and awards	personal characteristics
areas of knowledge	driver´s licence	honors, awards and prizes	personal data
areas of specialization	driver´s license	honors/activities	personal detail
army experience	driving licence	honours and awards	personal details
artistic competences	driving license	industrial training	personal dossier
artistic skills	education and positions	industry experience	personal experience
athletic involvement	education and qualification	instructional aide	personal information
awards & affiliations	education and training	int´l experience	personal interests
awards & recognitions	education qualifications	internship experience	personal objective
awards / achievements	educational background	invited lectures	personal profile
awards / fellowships	educational credentials	it skills	personal skills
awards and achievements	educational history	job description	personalities traits
awards and honors	educational information	job experience	personality traits
awards/memberships	educational qualification	job objective	primary education
		job profile	prior experience

Table A.4: Cue phrase list I. Containing of cue phrases discovered from the analysis.

A.5 Cue phrases II

activities & achievements
additional education & trainings
analytical / finance skills
practicum experience
pre-degree qualifications
peer reviewed papers
personal skills and abilities
personality traits / skills
proceedings and presentation
professional accomplishments
professional achievements
professional activities
professional affiliations
professional associations
professional background
professional certifications
professional credentials
professional development
professional experience
professional exposure
professional history
professional memberships
professional objective
professional profile
professional publications
professional qualifications
professional skills
professional statement
professional summary
professional training
proficiency in programming
programming languages
programming skill
programming skills
project experience
projects done
publications and papers
publications and software
published expertise
recent experience
recent trends
recent trends / research interest
references/portfolio
related achievement
related course projects
related courses
related experience
relevant experience
relevant skills
research interest
research activities
research and technology
research experience
research grants
research interests
research projects
research skills
reviewer for journals
rewards & recognition
rewards and recognition
scholarship received
school education
select computer skills
seminar attend
short biography
significant achievents
skills & abilities
skills & knowledge
skills & software
skills and attributes
skills and expertise
skills and knowledge
skills hci methods
skills module
skills profile
skills, expertise
skills/experience
skill set
social abilities
social competence
social media profiles
social skills
soft skills
software capabilities
software expertise
software skills
special skills
special skills and attributes
special training
specialized skills
sporting activities
strengths and key skills
strong skills
summary of qualification
summary of qualifications
summary profile
summer internship
summer internship program
supervision activities
teaching activities
teaching experience
teaching interests
technical background
technical competence
technical experience
technical knowledge
technical proficiencies
technical proficiency
technical profile
technical qualification
technical skill
technical skills
tertiary education
top skills
university service
university studies
values and skills
volunteer experience
volunteer experiences
volunteer work
volunteering activities
web portfolio
work experience
work history
work samples workshops & seminars
workshops and seminars
writing samples
academic background
academic experience
academic honors
academic training
activities and honors
activity and honor
additional experience
additional information
additional qualifications
additional skills
areas of experience
areas of expertise
areas of knowledge
army experience
athletic involvement
awards / achievements
awards and achievements
awards and honors
career goal
career objective
career objectives
career related experience
career related skills
career summary
civic activities
college activities
community involvement
computer knowledge
computer proficiency
computer skill
computer skills
conference presentations
course project experience
current research interests
curriculum vitae
driver license
driver's license
driving licence
driving license
extra-curricular activities
education and qualification
education and training
educational background
educational qualification
educational qualifications
educational training
employment history
employment objective
experience skills
freelance experience
graduation project done
grants and awards
grants and prizes
highlights of profession
highlights of qualification
honor and activity
honor and awards
honors and awards
honors, awards and prizes
honours and awards
industry experience
internship experience
it skills
job description
job experience
job objective
journal publications
key skills
key strengths

Table A.5: Cue phrase list II. Containing of cue phrases discovered from the analysis.

A.6 Cue phrases III

academic & professional qualification	personal data	social skills and abilities
academic/ non-academic achievements or prizes	personal detail	social skills and competences
accomplishments and honors	personal details	special award / honors / certification
achievements & projects done	personal experience	things that i want you to know
achievements and extra - curricular activities	personal information	community & networking activities
artistic skills and competences	personal interests	computer skills and abilities
peer reviewed papers and publications	personal objective	computer skills and competences
personal & professional profile	personal profile	documentary projects & collaborations
professional academic qualification	personal skills	tutor, instructional aide
professional experience, summary	professional accomplishments	teaching/supervision activities
professional experiences and achievements	professional activities	technical knowledge details
professional management skills	professional affiliations	technical skills and competences
professional memberships/affiliations	professional associations	technical experience
proffessional qualification / certification	professional background	technical knowledge
public relations experience	professional experience	technical proficiencies
highlights of professional skills	professional history	technical qualification
highlights of professional skills and accomplishments	professional memberships	technical skill
highlights of profession	professional objective	technical skills
highlights of qualification	professional profile	top skills
highlights of qualifications	professional publications	volunteer experience
hobbies, activities and interests	professional skills	volunteer work
knowledge and abilities	professional summary	web portfolio
knowledge, skills and abilities	professional training	work experience
journal and presentation	programming languages	work history
journal, proceedings & presentation	programming skill	writing samples
language and computer skills	publications and papers	employment record/university studies/teaching experience
language and skills	related achievement	executive summary & covering letter
language competencies and skills	related course projects	fellowships, awards, and grants
languages and skills	related courses	
language and skills	related experience	
language competencies and skills	relevant experience	
language skills	relevant skills	
languages and skills	research grants	
leaderships experience	research projects	
linguistic abilities	research skills	
languages, main areas of interest and other engagement	rewards & recognition	
links, interests and activities	rewards and recognition	
motivation and interests	seminar attend	
management skills	skills & knowledge	
military background	skills and accomplishment	
military experience	skills and knowledge	
mother tongue	social media profiles	
native language	software expertise	
native tongue	software skills	
other employment	special award / honors / certification	
other experience	special skills	
other skills	special training	
outside interests	specialized skills	
organisational skills	strengths and key skills	
organizational activities	summary of qualification	
organizational activities/ community involvement	summary of qualifications	
organizational skills and abilities	scientific community activities	
organizations and honors	skills and accomplishment	
organizations/ achievements	skills and accomplishments	
other professional activities	skills/interests software skills	
other skills and competences		

Table A.6: Cue phrase list III. Containing of cue phrases discovered from the analysis.

A.7 Topic continuation cue words/phrases

address	gender	residence
cell phone	marital status	responsibilities
civil status	mother name	responsible
city	mother language	role
country of residence	name	religion
date of birth	nationality	street
dissertation title	phone	sales
duties	postal code	sex
duty	position	support
name	projects	state
database	project title	spouse
father name	projects details	surename
founding member	periode	supervisor
funding		thesis title
funds		tel
fax		task
		telephone
		tools

Table A.7: List of cue phrases mined from the web. Source: "CV-headlines" (2015)

A.8 Action words

accelerated	billed	corrected	engineered	headed	maximized	selected		
accomplished	blazed	correlated	enhanced	helped	measured	separated		
accounted for	boosted	corresponded	enlarged	hired	mediated	proved	served	
accumulated	bought	corroborated	enlisted	hypothesized	merged	provided	serviced	
achieved	briefed	costed	ensured	identified	met	publicized	set	totaled
acquired	broadened	counseled	entertained	illustrated	minimized	published	shaped	toured
acted	budgeted	counted	established	imagined	mobilized	purchased	shared	traced
activated active	built	created	estimated	implemented	moderated	pursued	showed	tracked
adapted	calculated	critiqued	evaluated	impressed	modernized	qualified	simplified	traded
addressed	campaigned	crowned	examined	improved	modified	queried	simulated	trained
adjusted	captured	cultivated	exceeded	improvised	monitored	questioned	sketched	transcribed
administered	cataloged	cured	executed	incorporated	motivated	raised	slashed	transferred
advanced	caused	customized	expanded	increased	moved	ran	sold	transformed
advertised	centralized	cut	expedited	indexed	named	ranked	solidified	translated
advised	chaired	dealt	experienced	indoctrinated	navigated	rated	solved	transmitted
advocated	championed	decided	experimented	influenced	negated	reached	sorted	transported
affected	changed	decreased	explained	informed	negotiated	realigned	sought	traveled
aided	channeled	defined	explored	initiated	netted	realized	sparked	treated
alerted	charted	delegated	expressed	innovated	observed	reasoned	spearheaded	triggered
allocated	checked	delivered	extended	inspected	obtained	received	specialized	trimmed
amplified	clarified	demonstrated	extracted	inspired	opened	recognized	specified	tripled
analyzed	classified	described	fabricated	installed	operated	recommended	spoke	triumphed
answered	closed	designated	facilitated	instigated	optimized	reconciled	sponsored	troubleshot
anticipated	coached	designed	familiarized	instituted	orchestrated	recorded	staffed	turned
applied	co-directed	detected	fashioned	instructed	ordered	recruited	standardized	tutored
appointed	collaborated	determined	filed	insured	organized	redesigned	started	typed
appraised	collected	developed	filled	integrated	originated	reduced	steered	umpired
approved	co-managed	devised	finalized	interpreted	outlined	referred	stimulated	uncovered
arbitrated	combined	diagnosed	financed	interviewed	overhauled	registered	stored	understood
arranged	commanded	directed	fine-tuned	introduced	oversaw	regulated	streamlined	understudied
arraigned	commended	discovered	fixed	invented	participated	rehabilitated	strengthened	undertook
arrested	commented	dispatched	focused	inventoried	perceived	reinforced	stressed	underwent
articulated	communicated	dispensed	forecast	invested	performed	related	stretched	underwrote
ascertained	compared	displayed	forecasted	investigated	persuaded	remodeled	structured	unearthed
aspired	compiled	dissected	formed	involved	photographed	rendered	studied	unified
assembled	completed	distinguished	formulated	joined	piloted	reorganized	submitted	united
assessed	composed	distributed	fostered	judged	pinpointed	repaired	substituted	unraveled
assigned	computed	documented	found	justified	pioneered	replaced	succeeded	updated
assisted	conceived	doubled	founded	kept	placed	replied	suggested	upgraded
assumed	conceptualized	drafted	fulfilled	launched	played	reported	summarized	urged
assured	condensed	drove	functioned	lead	planned	represented	superseded	used
attained	conducted	earned	furnished	learned	predicted	reputed	supervised	utilized
attracted	conferred	economized	gained	leased	prepared	researched	supplemented	validated
audited	conserved	edited	gathered	lectured	presented	resolved	supplied	valued
authored	considered	educated	generated	led	presided	responded	supported	verbalized
automated	consolidated	effected	graded	liaised	prevented	restored	surpassed	verified
awarded	constructed	eliminated	graduated	licensed	printed	restructured	surveyed	visited
balanced	consulted	emphasized	granted	listed	prioritized	retrieved	synchronized	vitalized
	contacted	employed	grew	located	processed	revamped	synergized	volunteered
	contained	empowered	guided	logged	procured	reversed	systematized	waged
	contracted	enabled	halved	machined	produced	reviewed	tabulated	weighed
	contributed	enacted	handled	made	programmed	revised	tackled	widened
	controlled	encouraged	harmonized	magnified	prohibited	revitalized	targeted	won
	converted	ended	harnessed	maintained	projected	routed	taught	worked
	convicted	endorsed		managed	promoted	saved	terminated	wrote
	coordinated	energized	issued	marketed	proofread	scheduled	tested	
		enforced		mastered	proposed	screened	tightened	
		engaged		matched	protected	searched	took	
						secured		

Table A.8: List of mined action words. Source: "Action words" (2015).

A.9 Penn Part Of Speech tags

Tags	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Table A.9: Example of topic specific words.

CV topics				
contact information	personal data	honour and award	education	experience
address	name	diploma	bachelor	engineer
telephone	date of birth	honour	master	manager
mobile	nationality	honor	thesis	IEEE
email	place of birth	award	PhD	board
postal code	gender	scholarship	University	organisation
cell phone	citizenship	fellowship	gpa	
e-mail	country of residence	price	college	
residence	religion			
phone	father name			
street	marital status			
tel	place of birth			
email	sex			
	civil status			
	spouse			
	children			

Table A.10: Penn Part of Speech tags. Source: "Penn Treebank Project" (2015)